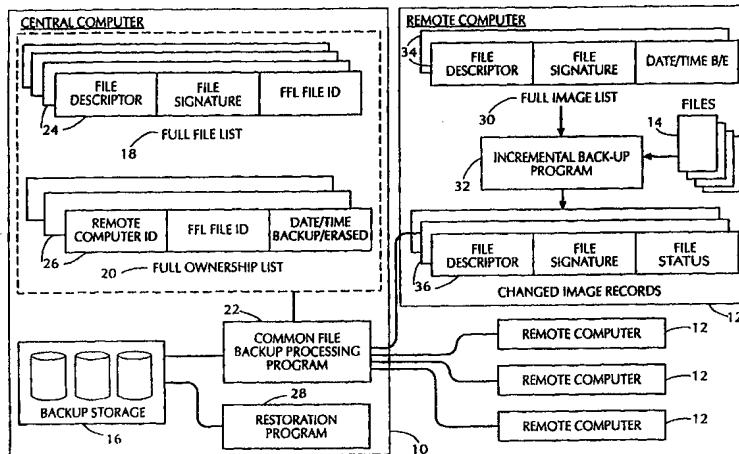


INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : G06F 11/14		A1	(11) International Publication Number: WO 99/09480 (43) International Publication Date: 25 February 1999 (25.02.99)
(21) International Application Number: PCT/IB98/01203		(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).	
(22) International Filing Date: 22 July 1998 (22.07.98)			
(30) Priority Data: 08/902,535 29 July 1997 (29.07.97) US			
(71) Applicant: TELEBACKUP SYSTEMS, INC. [CA/CA]; #400, 609 14th Street, Calgary, Alberta T2N 2A1 (CA).			
(72) Inventors: SWOVELAND, Cary; 310 2025 W. 42nd Avenue, Vancouver, British Columbia V6M 2B5 (CA). SOMERVILLE, Robert; 701 13104 Elbow Drive, Calgary, Alberta T2W 2P2 (CA).			
(74) Agents: BRETT, R., Allan et al.; Smart & Biggar, 900 – 55 Metcalfe Street, P.O. Box 2999, Station D, Ottawa, Ontario K1P 5Y6 (CA).			

(54) Title: METHOD AND SYSTEM FOR NONREDUNDANT BACKUP OF IDENTICAL FILES STORED ON REMOTE COMPUTERS



(57) Abstract

Method for backing-up files stored on remote computers to a central computer. The method includes storing a file list in a memory device accessible to the central computer. The file list contains multiple records which each correspond to a file stored at some time on one or more of the remote computers. Each record includes file identification data which identifies a respective file and includes a signature of the respective file. During a backup operation of a remote computer, file identification data is transmitted from the remote computer to the central computer. The central computer compares the file identification data, including the file signature, received from the remote computer with the file identification data of one or more records contained in the file list. If the file identification data received from the remote computer does not match the file identification data of any of the records contained in the file list, the central computer transmits a message to the remote computer instructing the remote computer to transmit the file to the central computer. The central computer further adds a record containing the file identification data to the file list.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

METHOD AND SYSTEM FOR NONREDUNDANT BACKUP OF IDENTICAL FILES STORED ON REMOTE COMPUTERS

COPYRIGHT NOTICE

5 A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent files or records, but otherwise reserves all copyright rights whatsoever.

10 BACKGROUND OF THE INVENTION

The invention disclosed herein relates generally to computer systems for backing up files. More particularly, the present invention relates to a method and system for backing up files stored on a large number of remote computers to a central computer while reducing storage and bandwidth requirements by minimizing or eliminating the 15 redundant storage of identical files stored on more than one of the remote computers.

Data stored in the hard drive files of personal computers and workstations are generally inadequately protected against loss. The danger of loss can arise when the hard drive becomes faulty through failure, destruction by fire, flood or other physical disasters as well as the data becoming corrupted, lost or, infected by a computer virus. Typical 20 backup methods for hard drives such as floppy disk media, attached tape drives, or duplicate hard drives require substantial manual initiation to set-up, define and operate,

storage media handling including off-site transportation for storage and retrieval, and dedication of the workstation users prime time to oversee the whole backup process.

Remote computer backup can be used to move copies of data files from one computer over a transmission medium such as a telephone line to an offsite computer and 5 storage medium. A remote backup guards against many of the problems which give rise to loss of data in a computer's hard drive. The protection is provided by the presence of a number of previous copies of the files which can be restored to the workstation based on the date of the desired copy.

Attempts to unburden the workstation user through automatic on-line processes 10 have consistently run into communications issues, the prime one being the amount of stable connected time (bandwidth) to achieve the initial complete backup and subsequent incremental backups on a regular cycle. Additionally, the amount of storage required has had a significant impact on media selection, number of backup sets maintained, and cost of maintenance.

15 Existing remote backup systems have not been able to effectively reduce the communications time required for the first full backup of a new workstation and subsequent backups, and the amount of storage required to perform the backup. There is thus a need for a backup process which is designed to solve the problems associated with backups of remote computers and to reduce the bandwidth and storage requirements 20 associated with existing backup processes.

BRIEF SUMMARY OF THE INVENTION

It is an object of the present invention to provide an improved backup method and system which solves the problems set forth above associated with existing backup systems.

5 It is another object of the present invention to reduce the bandwidth required for the backup of files stored on a remote computer to a central computer.

It is another object of the present invention to reduce the storage requirements in a central computer backing up a number of remote computers.

10 It is another object of the present invention to minimize or eliminate the redundant storage on a central computer of identical files backed up from a number of remote computers.

The above and other objects are achieved by a method for backing up files stored on remote computers to a central computer, the remote computers being connectable to the central computer. The method involves storing a file list in a memory 15 device accessible to the central computer, the file list containing a plurality of records each record corresponding to a file stored at some time on one or more of the remote computers, and each record comprising file identification data identifying a respective file and including a signature of the respective file. During a backup operation of a first remote computer, first file identification data is transmitted from the first remote 20 computer and received at the central computer. The first file identification data identifies

a first file stored at some time on the remote computer and includes a signature of that file.

The central computer compares the first file identification data including the first file signature received from the first remote computer with the file identification data of 5 one or more records contained in the file list. If the first file identification data received from the first remote computer does not match the file identification data of any of the records contained in the file list, the central computer transmits a message to the first remote computer instructing the first remote computer to transmit the first file to the central computer. The central computer further adds a record containing the first file 10 identification data to the file list.

BRIEF DESCRIPTION OF THE DRAWINGS

The invention is illustrated in the figures of the accompanying drawings which are meant to be exemplary and not limiting, in which like references refer to like or corresponding parts, and in which:

15 Fig. 1 is a block diagram showing the backup system of one embodiment of the present invention;

Figs. 2A-2C show the sequence of messages transmitted between the central computer and remote computer during a backup operation in accordance with one embodiment of the present invention;

Figs. 3A-6 contain flow charts showing the method of backing up a remote computer onto the central computer of the system shown in Fig. 1 in accordance with one embodiment of the present invention, of which:

5 Figs. 3A-3B show the process performed by the remote computer of creating changed image records for transmission to the central computer;

Figs. 4A-4B show the process performed by the central computer of processing the changed image records received from the remote computer as a result of the process shown in Figs 3A-3B and creating modified changed image records for transmission to the remote computer;

10 Figs. 5A-5B show the process performed by the remote computer of processing the modified changed image records received from the central computer as a result of the process shown in Figs. 4A-4B and selecting files to be transmitted to the central computer for backing up; and

Fig. 6 shows the process performed by the central computer of storing the 15 files received from the remote computer as a result of the process shown in Figs. 5A-5B;

Figs. 7A-7C contain a flow chart showing the process of restoring files to the remote computer; and

Figs. 8-15 are screen displays displayed by one version of a software program implementing an embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The method and system of the present invention will be described herein with reference to the system diagrams shown in Figs. 1 and 2, the flow charts shown in Figs. 3A-7C, and the screen displays shown in Figs. 8-15. One skilled in the art will recognize 5 that these drawings illustrate one manner of implementing the present invention and that many variations are possible within the scope of the present invention.

As shown in Fig. 1, the system of one embodiment contains a central computer 10 and a number of remote computers 12 connected to the central computer 10. The central computer 10 may be a single computer or a group of networked computers, operated as 10 the backup file repository. The remote computers 12 may be personal computers, workstations attached over a network, network servers, or other types of computers or computer systems. The remote computers 12 may be connected to the central computer by virtue of any conventional communications link, including copper wire network, coaxial cable network, satellite network, fiber optics network, or microwave network. 15 Information may also be transmitted between the central and remote computers by physical media transfer such as by disk, tape or other information storage device. As such, the remote computers 12 are not necessarily distanced from the central computer 10 by any specific amount.

The remote computers 12 (a representative one being shown schematically in Fig. 20 1) contain storage devices such as direct access storage devices on which are stored a number of files 14. In accordance with well-known practice, each file 14 has associated

therewith a file descriptor consisting of a record containing information about the file (other than the contents of the file itself) that is stored by the disk operating system of the remote computer 12. File descriptors includes the name of the file, the location of the file in the file system (such as the drive and directory), the size of the file, and its date/time 5 stamp. The size of the file is usually represented by the number of bytes in the file. The date/time stamp usually represents the date and time the file was last changed, regardless of whether the file was last changed while on the remote computer 12 or last changed on another computer and copied to the remote computer 12. If the file was never changed (after it was created) on the remote computer 12 or some other computer, the date/time 10 stamp is the date and time the file was created. Depending upon the operating system, the file descriptor may also include the date and time the file was created, the type of file (e.g., read-only, system, hidden), whether an archive bit is set or not, and extended file attributes as known to those of ordinary skill in the art.

For purposes of the system of the preferred embodiments, two files 14 are 15 considered to be different when one file does not contain exactly the same information as the other file. Because the focus is on the content of the files, two files with different file names or other identifying characteristics may not differ in this sense. To distinguish between files, the system computes a signature for each file 14. In preferred embodiments, a file's signature is the numeric value of a mathematical function that is 20 applied to all the bytes in the file. This function belongs to a class of polynomial functions called "cyclic redundancy checks," which are well-known functions developed

to detect errors in modem transmissions. The particular polynomial function used in one preferred embodiment produces two integers -- one four byte, one two byte -- which combined represent any value between zero and approximately one trillion. As one skilled in the art will recognize, however, other mathematical functions could be used for 5 the purpose of computing a file's signature, e.g., a file digest, exclusive-oring, etc. The particular function used for the signature represents a balance between accuracy and speed as understood by those of skill in the art. In accordance with the invention, two files are considered different by virtue of their signatures being different and two files are considered identical by virtue of at least having identical signatures and, in preferred 10 embodiments, by virtue of having identical sizes and signatures.

The central computer 10 has storage devices 16 for storing backup copies of files 14 stored on the remote computer 12. The storage devices 16 may be any conventional storage devices including magnetic storage such as hard disks and optical storage such as writable CD-ROMs. In some embodiments, the storage devices 16 consist of two or more 15 different types of devices having differing capabilities to serve as primary and secondary storage devices. For example, the storage devices 16 of one embodiment are high capacity (4.25 gigabyte), SCSI high speed on-line hard drive storage disks serving as primary storage and high capacity (8 gigabyte), 4 mm DAT tape drives for disaster recovery. The capacity and number of such storage devices depends upon the needs of 20 the system as would be understood by those of skill in the art.

In accordance with one aspect of the present invention, the central computer 10 serves as a “common file archive” for all the files 14 stored on all remote computers 12 connected or connectable to the central computer 10. That is, the central computer maintains a virtual repository of files which are the same for two or more remote 5 computers. These files are not physically transferred from the second and subsequent remote computers which contain them. This common file archive ensures that the only files physically transferred from a remote computer are truly unique to that computer. Hence, bandwidth capacity to accommodate the backup process is minimized as well as the storage required for the backup set.

10 To implement this common file archive, the central computer 10 contains a full file list 18 (“FFL”), a file ownership list (“FOL”) 20, and a common file backup processing program 22. The FFL 18 is a set of records 24 each corresponding to a file 14 that is or was located on one or more remote computers 12. Each FFL record 24 contains the file descriptor for the corresponding file 14, a signature of the file 14, a code identifying the 15 remote computer 12 operating system from which the file 14 was backed up, and a FFL File ID that is unique to each FFL record which represents a file stored on one or more of the remote computers 12.

The FOL 20 is a set of records 26 each of which corresponds to a particular remote computer and either a file that is or was located on that computer, or a record of a file on 20 the remote computer that has been erased or overwritten. Each record contains a unique identifier for the remote computer, which is assigned by the central computer for each

remote computer to which it is or may be connected, the FFL File ID for the associated record in the FFL, and the date and time the file was backed up on the remote computer, or was recognized as having been newly-erased, as explained further below.

In accordance with the invention, and as explained in greater detail below, the FFL 5 18 is used by the central computer 10 to determine whether a file 14 to be backed up by a given remote computer 12 is already backed up in the backup storage 16 from another remote computer 12, and the FOL 20 is used to keep track of which of the files 14 identified by the FFL records 24 are stored on each remote computer 12 at a given time. Both the FFL 18 and FOL 20 may be stored in the central computer 10 as a set of records 10 or as entries in a one or more databases or tables.

The common file backup processing program 22 manages the backup processes performed on the central computer 10, as explained below. The central computer also has a restoration program or routine 28 which manages the process of restoring a given remote computer 12 following a hard drive or other failure or data loss.

15 In preferred embodiments, each remote computer 12 has a full image (“FI”) list 30 and an incremental backup program 32. The FI 30 is a set of records 34 maintained on each remote computer 12 that stores information regarding files which have previously been backed up on the central computer 10 from that remote computer 12. The FI contains all information contained in the records 24, 26 in the FFL 18 and FOL 20, 20 respectively, relating to that particular remote computer 12, except for the FFL File ID. The FI 30 is used by the incremental backup program 32 during a backup operation

performed at the remote computer 12, as described further below, to determine whether each file 14 then stored on the remote computer 12 has already been backed up on the central computer 10, and whether any files previously backed up to the central computer 10 from this remote computer 12 have since been erased from the remote computer 12.

5 The product of the backup program 32 is a set of changed image ("CI") records 36. The CI records 36 represent the changes to the files 14 stored on the remote computer 12 since a prior backup operation to the central computer 10. Each CI record contains the file descriptor and file signature for each file whose status has changed since the prior backup, and a status field indicating whether the specified file is new, changed, or newly-
10 erased. These file status options, as used in the preferred embodiments, have the following meanings:

(i) a file 14 is considered "new" to a given remote computer 12 if it was not present in the same location in the remote computer's file system when the remote computer 12 was last backed up. For most computer operating systems, this means that
15 the file 14 was not present in the same directory on the same logical drive when the remote computer 12 was last backed up;

(ii) a file 14 is considered "changed" to a given remote computer 12 if it differs from, but has the same name, as a file that was present in the same location in the remote computer's file system when the remote computer 12 was last backed up. For most
20 remote computer operating systems, this means that there was a different file by the same

name in the same directory on the same logical drive when the remote computer was last backed up; and

(iii) a file 14 is considered "newly-erased" to a given remote computer 12 if it was present when the remote computer 12 was last backed up, but is no longer present in the 5 same location in the remote computer's file system. For most remote computer operating systems, this means that the file that was present when the computer was last backed up is no longer present in the same directory on the same logical drive.

In preferred embodiments, a single byte is used for the file status field, using the representations "N" for "new," "C" for "changed," and "E" for "erased."

10 With respect to these status types, a file 14 on the remote computer 12 is considered to be "backed up" when an exact copy of the file 14 and the file descriptor is stored on the central computer 10, and a record of the existence of the file 14 and file descriptor on the remote computer 12 has been created and stored on the central computer 10.

15 An overview of the sequence of communication between the central computer 10 and a remote computer 12 during a backup operation is shown in Figs. 2A-2C. Each step in the process is described in more detail below. As shown in Fig. 2A, the remote computer 12 produces a set of CI records 36 and transmits these records to the central computer 10.

20 The central computer 10 processes the CI records 36 and produces modified CI records 36P and 36L which are transmitted to the remote computer 12 as shown in Fig.

2B. Some modified records 36P contain a field with an instruction to the remote computer 12 to perform a physical transfer of the file 14 identified by the CI record 36. A physical backup is the process of transmitting a copy of a file 14, together with its file descriptor, from the remote computer 12 to the central computer 10, where the file is stored, and creating and storing on the central computer 10 a record of the existence of this file 14 and its file descriptor. Other modified CI records 36L contain a field with an instruction to the remote computer 12 to perform a logical transfer of the file 14 identified in the CI record. A logical backup is the operation of transmitting a copy of a file descriptor from the remote computer 12 to the central computer 10 (but not the file itself), where the file descriptor is stored, and creating and storing a record of the existence of the file descriptor and associated file on the remote computer 12.

As shown in Fig. 2C, the remote computer 12 responds to the modified CI records by transmitting the files 14 identified in the modified CI records 36P, which contained an instruction to perform a physical transfer, to the central computer 10. To reduce bandwidth and provide for security and privacy, the files 14 are compressed and encrypted before transmission.

Each step in this backup process is described in more detail with reference to the flow charts in Figs. 3A-6.

The process of creating the CI records is shown in Figs. 3A-3B. The backup process is started at the remote computer, step 50. The process can be started in one of two ways:

(i) the operator of the remote computer 12 can cause the backup program to be executed, in the same way the execution of other application programs are initiated; or

(ii) a program running in the background on the remote computer 12 can initiate the backup operation at a particular time-of-day which may be scheduled and modified by

5 the operator of the remote computer 12.

Each file 14 stored on the remote computer 12 is then retrieved, step 52, and a signature is computed for the file, step 54. If the remote computer 12 has not previously been backed up to the central computer, all the files 14 would be new, and a CI record is created for each file containing the respective file's file descriptor, file signature and a

10 status field set to "new."

If a prior backup operation has been performed for the remote computer 12, the file descriptor for each file 14 is compared to the file descriptors in the FI records, step 56. In particular, the file name and file location parts of the file descriptor are used in this comparison. If no file name and location in the FI records' file descriptors match a given

15 file's file name and location, step 58, the file status is set to "new," step 60, and a CI record is created with the file's file descriptor, file signature, and file status of "new," step 62 (Fig. 3B).

If the file name and location for a given file 14 match a file name and location of a FI record, step 58, then it is understood that the file 14 or a version thereof has previously

20 been backed up to the central computer 10. A flag associated with the FI record is set, step 64, indicating that a file corresponding to this FI record exists and has not been

erased from the remote computer 12. Tests are then performed to determine whether the file 14 has changed since the last backup. In some embodiments, this determination is made through use of the size and signature of the file. That is, the size of the file is compared to the file size stored in the FI record's file descriptor, step 66. If the file sizes 5 do not match, then the file 14 has changed since the last backup. If the file sizes match, the file signature of the file is compared to the file signature stored in the FI record, step 70. If the file signatures are different, then the file has changed. In either case of non-matching values, the file status is set as "changed," step 68, and a CI record is created with the file descriptor, file signature, and file status of "changed," step 62.

10 If the file signatures match, then the file is considered not to have changed since the previous backup operation. As a result, no CI record is created for this file. If there are any more files, step 72, the next file is processed in the same manner.

 In alternative embodiments, only the file sizes are compared, thus saving time by avoiding the need to compute file signatures for all files stored on the remote computer.

15 In further embodiments, file signatures are only computed and compared for those files having a file size which matches a file size stored in a FI record. The remote computer user may be given the option of using file signatures to achieve greater accuracy in the comparison process.

 When all remote computer files 14 have been so processed, the FI records are 20 checked, step 74, to determine whether any FI records were not accessed, thus showing that the files represented by those FI records have been erased. For each FI record, the FI

record accessed flag is checked, step 76. If the flag is not set, then the file status is set to “erased,” step 78, and a CI record is created containing the file descriptor and file signature from the FI record and a file status of “erased,” step 80. If there are any more FI records to check, step 82, the same process is performed on each.

5 When this process is complete, a set of CI records will have been created representing files which are new, changed and erased since the prior backup, if there had been one. The date and time at which the backup operation was initiated on the remote computer is attached to each CI record. The CI records 36 are transmitted to the central computer 10, step 84.

10 One skilled in the art will recognize that the process just described of identifying files which are new, changed or erased may be performed entirely on the central computer 10 by transmitting the file descriptors and file signatures for all files 14 stored on the remote computer 12 to the central computer 10. However, performing this process on the remote computer 12 reduces the amount of data needed to be transmitted to the 15 central computer 10, thus reducing the bandwidth required in the system. Furthermore, other means could be used to determine whether a file on the remote computer 12 has changed, such as by examination of the file’s “archive bit,” as understood by those of skill in the art, if this information is included in the file’s file descriptor.

20 The central computer 10 processes the CI records 36 in accordance with the procedure shown in Figs. 4A-4B. The central computer 10 receives the CI records, step 100, and determines the remote computer ID, step 102. The remote computer ID may be

found in a look up table or similar data structure stored on the central computer 10, or may be attached as one of the first parameters sent by each remote computer 12 with a set of CI records. The central computer opens the FFL and FOL, step 104, which will be used to determine whether any of the files identified by the CI records were already 5 obtained from another remote computer 12 and are already stored in the backup storage 16.

Each CI record is opened, step 106, and the status field checked, step 108. If the file status is set to "E" for "erased" (meaning that the file identified by the CI record is stored in the backup storage), the FFL is checked to find the FFL record which 10 corresponds to the CI record. In one embodiment, this is accomplished by retrieving each FFL record, step 110, and comparing the FFL record's file descriptor to the file descriptor in the CI record, step 112. If the file descriptors match, the file signatures from the FFL record and CI record are compared, step 114. If the file signatures match, then the current FFL record represents the file which has been erased on the remote computer 12, and 15 another record is added to the FFL which is a duplicate of the current FFL record, step 116. The duplicate record is inserted into or indexed in the FFL in a position immediately following the current FFL record. This duplicate, "deletion" record is used during the restoration process in a manner described below. An action code, which will be stored in the FOL, is set to "newly erased," step 118.

20 In preferred embodiments, the original FFL record is not immediately deleted from the FFL, nor is the compressed file stored in backup storage 16 immediately deleted.

Rather, a counter is associated with each FFL record which counts the number of remote computers storing the file identified by the FFL record. The counter is incremented (increased by one) each time a backup operation is performed as described herein in which a remote computer transmits a CI record indicating that it is storing a “new” file 5 corresponding to the FFL record. The counter is decremented (decreased by one) each time a remote computer sends a CI record indicating that the corresponding file has been erased or changed from a remote computer. If a FFL record’s counter is non-zero, the FFL record and corresponding compressed file are retained by the central computer. When a counter reaches zero, the FFL record and corresponding compressed file stored in 10 backup storage are either deleted immediately, at a scheduled time set by a system administrator with other zero-counter FFL records, or when the FFL record has reached a certain age as set by the system administrator and determined by reference to the corresponding FOL record.

If the file status field is not equal to “E” for “erased,” step 108, then the CI record 15 represents a new or changed file. The FFL is then checked to determine whether the identical file is already stored in the backup storage 16. In particular, each FFL record is retrieved, step 120, and the file size and signature from the FFL record is compared to the file size and signature from the CI record, step 122. If the file sizes and signatures match, the file descriptors from the FFL record and CI record are compared, step 124. If the file 20 descriptors match, the file identified by the CI record is deemed to be identical to a file already stored in the backup storage 16. Thus, a copy of the file 14 need not be

transmitted from the remote computer 12 to the central computer 10. To indicate this, a “logical transfer” code is added to the CI record, step 126, thus creating a modified CI record which will be transmitted back to the remote computer 12. In addition, the action code, used in the FOL, is set to “file backed up,” step 128.

5 If either the file size/signature or file descriptors from the CI record do not match any FFL record, the file 14 is deemed to be new, i.e., not already stored in the backup storage 16. A “physical transfer” code is added to the CI record, step 130, to serve as an instruction to the remote computer 12 to physically transfer the file to the central computer 10 for backup.

10 One skilled in the art will recognize that any suitable search routine may be used for locating FFL records having values that match those in a given CI record, including sorting, queries, binary tree searches, etc. Furthermore, one skilled in the art will recognize that different parameters from the CI records and FFL records may be compared, or may be compared in a different sequence, provided that the file signature 15 from a CI record is compared to the file signature in the FFL record before a determination is made that the files are identical.

For erased and changed files, a record is added to the FOL, step 132, containing the remote computer ID, FFL File ID of the current FFL record (which matched the CI record), and action code and the current date and time. This FOL record is used to 20 identify the state of files on the remote computer as of a given date and time, and is used during the restoration process. In some embodiments, a FFL and FOL record is also

created at this point for new files, but a flag is set for each record to indicate that the file has not yet been received. When the file is received from the remote computer, the flag is reset to zero. In other embodiments, the FFL and FOL records for new files are not created until the file is actually received from the remote computer, as shown in Fig. 6
5 and discussed below.

If all CI records have been similarly processed, step 134, the modified CI records are transmitted from the central computer 10 to the remote computer 12, step 136. The modified records instruct the remote computer 12 what action to take with respect to each file identified by a CI record.

10 The process performed on the remote computer 12 is shown in Figs. 5A-5B. The remote computer 12 receives the modified CI records, step 150, and opens each, step 152. If the modified CI record contains a record code of “physical transfer,” step 154, the remote computer 12 retrieves the file 14 having a file descriptor equal to the file descriptor in the modified CI record, step 156. The remote computer 12 compresses the
15 file, step 158, and encrypts it if desired. The compressed file 14 is transmitted to the central computer 10 along with the file descriptor and signature, step 160. An action code to be placed in a new FI record is set to “backed up,” step 164.

If the modified CI record contains a “logical transfer” code, step 162, then the remote computer 12 does not send the file to the central computer, but simply sets the
20 action code to “backed up,” step 164. If the modified CI record contains neither a

“physical transfer” or “logical transfer” code, then the file identified by the CI record was erased, and the action code is set to “erased,” step 166.

For each modified CI record, a FI record is added, step 170, containing the file descriptor and file signature from the modified CI record, the action code, and the date 5 and time the backup was initiated, which was attached to the CI record as explained above. New FI records with action codes of “erased” are preferably stored or indexed to immediately follow the corresponding FI record with the same file descriptor and signature. If there are no other modified CI records, step 172 (Fig. 5B), a message is transmitted to the central computer 10 that “backup is complete,” step 174.

10 The central computer processes the compressed files in accordance with the process shown in Fig. 6. For each file, the central computer receives the file descriptor, file signature and compressed (and possibly encrypted) file 14, step 190. The compressed file 14 is stored in backup storage 16, step 192, and new FFL and FOL records are created by assigning a new FFL File ID, step 194, adding a record to the FFL containing the file 15 descriptor, file signature and FFL File ID, step 196, and adding a record to the FOL containing the remote computer ID, FFL File ID, an action code of “File backed up,” and the date/time the backup operation was initiated (as attached to the CI record as explained above), step 198.

The process in accordance with one embodiment of the invention of restoring lost 20 files to the remote computer is illustrated by the flow chart of Figs. 7A-7C. A remote computer user initiates a restoration process, step 210, by either running a restoration

program on the remote computer, by connecting to the central computer and accessing a restoration program, or by calling the central computer administrator.

If the process is initiated on the remote computer, the restoration program searches the FI records to retrieve all date/time stamps in which backups were performed, step 212.

- 5 The backup dates are displayed to the user, step 214, and the user selects the desired date from the list, step 216. The restoration program on the remote computer then retrieves each FI record, step 218, and checks whether the date/time stamp predates (meaning, as used herein, is before or equal to) the selected backup date, step 220. If it does, the program checks, step 222, for the presence of any deletion records which, in the preferred embodiment, are FI records having duplicate file descriptors and file signatures as the current FI record but an action code of "erased." In some embodiments, these deletion records are located or indexed in the FI list to immediately follow the original FI record, as explained above. If any deletion record exists and has a date/time stamp which predates the user-selected backup date, step 222, a FI backup flag is not set, and, if there 10 are more FI records to process, step 226, the next FI record is retrieved, step 218.
- 15

- 20 If there is no deletion record for the current FI record, or if any deletion record does not predate the backup date, a FI backup flag is set, step 224, indicating that the FI record corresponds to a file to be restored. If there are more FI records to process, step 226, the next FI record is retrieved, step 218 and the process repeated until all FI records have been checked.

If the restoration process is performed on the central computer, either by the remote computer user remotely accessing the central computer restoration program or by the central computer administrator, a similar checking process is performed with records in the FOL and FFL. All FOL records with a remote computer ID corresponding to the 5 particular remote computer are retrieved and the date/time stamps checked to obtain a list of dates. The user selects the desired backup date, and FOL records which predate that backup date are retrieved. The FFL records having FFL File IDs matching those in the retrieved FOL records are retrieved and, accounting for deletion records, are used to obtain a list of files.

10 If the remote computer is used and the files are to be transmitted on-line or over a network or other communication link, step 230 (Fig. 7B), a list of the files corresponding to the FI records with backup flags set is displayed to the user, step 232. The user selects the desired files to be restored, step 234, and the selected FI records are stored in a restoration requested database, step 236. The database is transmitted to the central 15 computer, step 238, which receives the database, opens it, step 240, and retrieves the compressed files from backup storage 16, step 242. The central computer transmits the retrieved compressed files to the remote computer, step 244, and the files are decompressed at the remote computer, step 246 and placed in their proper location (drive and directory) on the remote computer. The restoration process is thus complete.

20 If the files are not to be transmitted from the central computer (Fig. 7C), the FI records, either transmitted from the remote computer or obtained from the FOL and FFL

as explained above, are stored in a restoration database, step 250, and the central computer retrieves the compressed files corresponding to the FI records from backup storage 16, step 252. The retrieved compressed files are stored on a high capacity storage medium such as a CD-ROM, step 256, and the storage medium is sent to the remote 5 computer user, step 256. The files from the storage medium are then loaded into the remote computer, step 258, and decompressed, step 260.

Figs. 8-15 contain screen displays used in one implementation of the present invention in a software program developed by the assignee of the present application, Telebackup Systems Inc. of Calgary, Alberta. The program is designed to operate with 10 the Windows95 operating system available from Microsoft Corporation of Redmond, Washington. Fig. 8 shows a configuration screen which allows for the entry of a user name and organization and for the selection of a particular modem and modem parameters. Fig. 9 shows a billing screen by which the central computer administrator tracks users for billing purposes.

15 Fig. 10 shows an installation screen in which a user is assigned a password. The user must employ the password to gain access to its files backed up on the central computer. Fig. 11 shows a backup scheduling screen in which the remote computer user can specify times in which backup is to be performed and times when backup should not be performed. Fig. 12 shows a backup configuration dialog box in which the user can 20 specify which drives to backup. Fig. 13 shows a file exclusion dialog box in which the user can specify which files should not be backed up.

Fig. 14 shows a file restoration dialog box in which the user is presented with the dates available for backup and, once a backup date is selected, the files available to be restored for that date. The user can also select whether the files should be sent by modem transmission or stored on a CD-ROM. Fig. 15 shows a restoration status dialog box 5 showing the progress of an on-line file restoration.

While the invention has been described and illustrated in connection with preferred embodiments, many variations and modifications as will be evident to those skilled in this art may be made without departing from the spirit and scope of the invention. The invention is thus not to be limited to the precise details of methodology or construction 10 set forth above as such variations and modification are intended to be included within the scope of the invention.

WHAT IS CLAIMED IS:

1. In a data processing system comprising a plurality of remote computers which are connectable to a central computer, a method for backing up files stored on the remote computers to the central computer, comprising:
 - 5 storing a file list in a memory device accessible to the central computer, the file list containing a plurality of records each record corresponding to a file stored at some time on one or more of the remote computers, each record comprising file identification data identifying a respective file and including a signature of the respective file;
 - 10 receiving at the central computer from a first remote computer first file identification data which identifies a first file stored at some time on the remote computer and which includes a signature of the first file;
 - 15 comparing the first file identification data including the first file signature received from the first remote computer with the file identification data of one or more records contained in the file list; and
 - 20 if the first file identification data received from the first remote computer does not match the file identification data of any of the records contained in the file list, transmitting a message from the central computer to the first remote computer instructing the first remote computer to transmit the first file to the central computer.
2. The method of claim 1 comprising, if the first file identification data received from the first remote computer does not match the file identification data of any of the

records contained in the file list, adding a record containing the first file identification data to the file list.

3. The method of claim 1 comprising, for each of a plurality of files stored at some time at the first remote computer, determining whether the file is:

5 a new file which has not previously been backed up from the remote computer to the central computer;

a changed file for which an earlier version thereof has previously been backed up from the remote computer to the central computer; or

a deleted file which has previously been backed up from the remote computer to

10 the central computer and which has since been deleted from the first remote computer, and wherein the first file identification data received at the central computer identifies a new, changed, or deleted file.

4. The method of claim 3 comprising receiving at the central computer status data associated with the first file identification data, the status data indicating whether the first

15 file identification data identifies a new, changed, or deleted first remote computer file.

5. The method of claim 4 comprising, if the status data indicates that the first file identification data identifies a new or changed file, and if the first file identification data received from the first remote computer matches file identification data of a record contained in the file list, storing an indication that the first file identified by the first file

20 identification data is stored on the first remote computer.

6. The method of claim 5 comprising storing a file ownership list in a memory device accessible to the central computer, the file ownership list containing a plurality of records each of which corresponds to a file stored on one remote computer, each record containing a remote computer identifier uniquely identifying one of the plurality of 5 remote computers and a pointer to a file list record, and wherein the step of storing an indication that the first file identified by the first file identification data is stored on the first remote computer comprises adding a record to the file ownership list containing an identifier for the first remote computer and a pointer to the matching file list record.

7. The method of claim 6 wherein each record in the file list has a unique file list 10 identifier, and wherein the step of adding a record to the file ownership list containing a pointer comprises adding a record containing the file list identifier of the matching file list record.

8. The method of claim 5 comprising, if the status data indicates that the first file 15 identification data identifies a new or changed file, and if the first file identification data received from the first remote computer matches the file identification data of a record contained in the file list, transmitting a message from the central computer to the first computer indicating that the first file need not be transmitted to the central computer.

9. The method of claim 4 comprising, if the status data indicates that the first file 20 identification data identifies a deleted file, storing an indication that the first file identified by the first file identification data has been deleted from the first remote computer.

10. The method of claim 9 comprising, if the status data indicates that the first file identification data identifies a deleted file, the step of locating a record in the file list containing file identification data which matches the first file identification data, and wherein the step of storing an indication that the first file identified by the first file identification data has been deleted from the first remote computer comprises adding a record to the file list which is identical to the located matching file list record.

5 11. The method of claim 1 comprising:

10 storing on the first remote computer an image file containing a plurality of records each of which corresponds to a file stored at some time on the first remote computer and backed up on the central computer, each image file record containing (i) identification data identifying a respective file including a signature of such file and (ii) an indicator of whether the respective file has previously been backed up on the central computer; for each of a plurality of files stored on the remote computer, comparing respective file identification data for that file with the file identification data of one or more image 15 file records.

12. The method of claim 1 wherein the step of receiving first file identification data at the central computer comprises receiving first file identification data which identifies a plurality of files stored at some time on the first remote computer.

13. The method of claim 12 comprising, for each of the plurality of files identified 20 by the first file identification data, determining whether the file is:

a new file which has not previously been transmitted from the first remote computer to the central computer;

a changed file for which an earlier version thereof has previously been transmitted from the first remote computer backed up to the central computer; or

5 a deleted file which has previously been transmitted to the central computer from the first remote computer and which has since been deleted from the first remote computer; and

performing the step of comparing first file identification data to file identification data contained in the file list only with respect to new, changed, or deleted files.

10 14. The method of claim 1 comprising storing a file ownership list in a memory device accessible to the central computer, the file ownership list containing a plurality of records each of which corresponds to a file stored on one of the plurality of remote computers, each record containing a remote computer identifier uniquely identifying one of the plurality of remote computers and a pointer to a given file list record.

15 15. The method of claim 14 comprising, upon a request to restore all or some of the files stored on the first remote computer, the steps of:

retrieving records from the file ownership list having remote computer identifiers corresponding to the first remote computer;

20 retrieving file list records pointed to by the pointers in the retrieved file ownership file records; and

transmitting the files identified by the file identification data contained in the retrieved file list records from the central computer to the first remote computer.

16. The method of claim 15 wherein the file ownership list records each contain a stamp indicating the date or time a file identified by the file identification data contained
5 in the file list record pointed to by the pointer was received by the central computer or was indicated as deleted from the identified remote computer, and wherein only some of the files stored on the first remote computer are to be restored, comprising specifying a date or time of the files to be restored and selecting the files to be restored from among the file ownership list records.

10 17. The method of claim 1 comprising, if the first file identification data received from the first remote computer matches the file identification data of at least one of the records contained in the file list, transmitting a message from the central computer to the first remote computer indicating that the first file need not be transmitted to the central computer.

15 18. The method of claim 1 comprising:
storing a counter associated with each record in the file list; and
if the first file identification data received from the first remote computer matches the file identification data of at least one of the records contained in the file list,
incrementing the counter associated with the record containing the matching file
20 identification data.

1/20

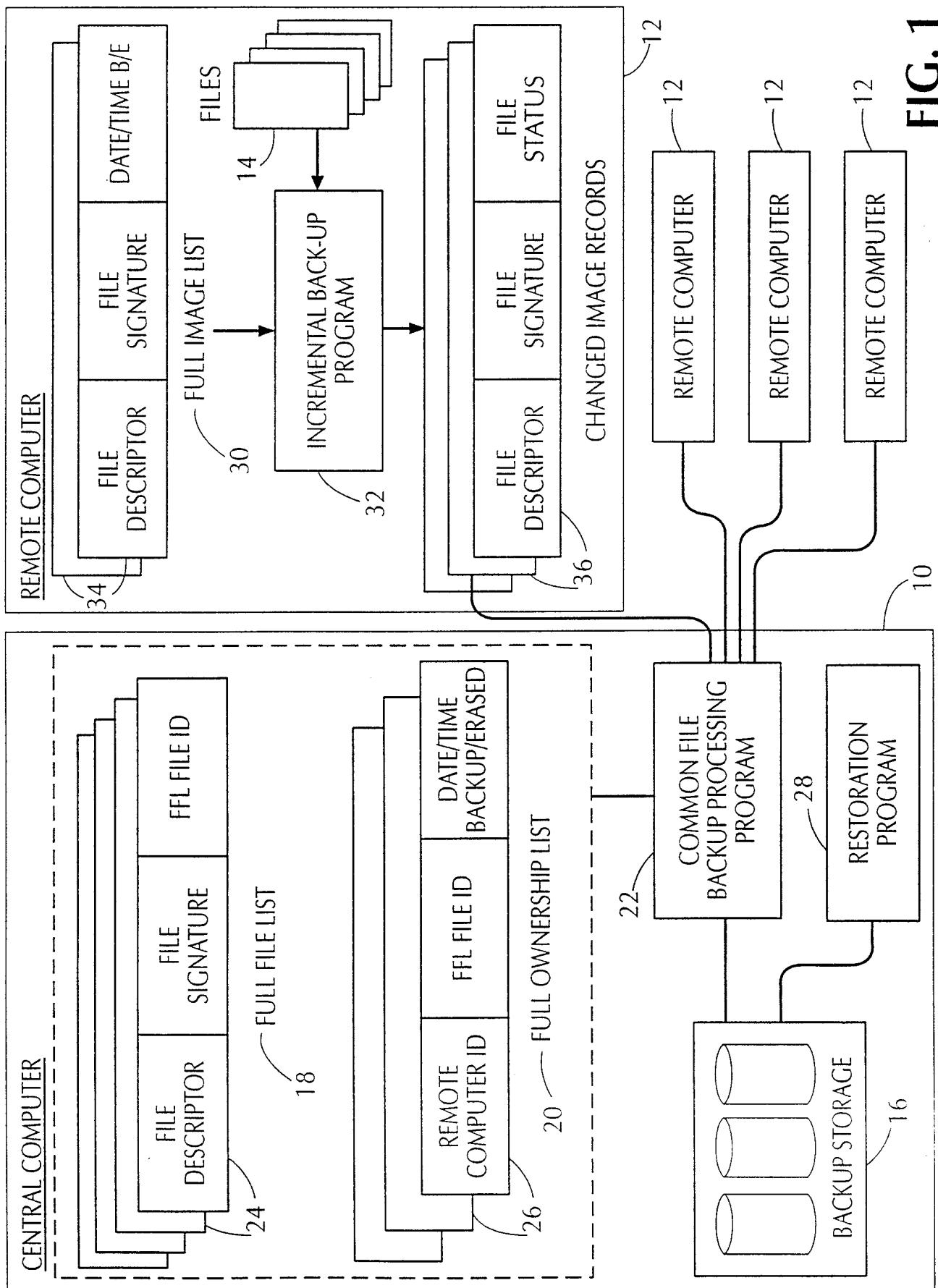
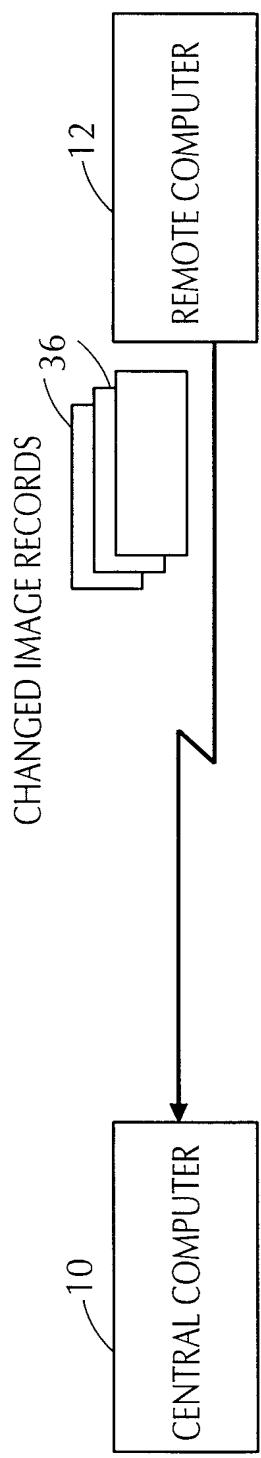
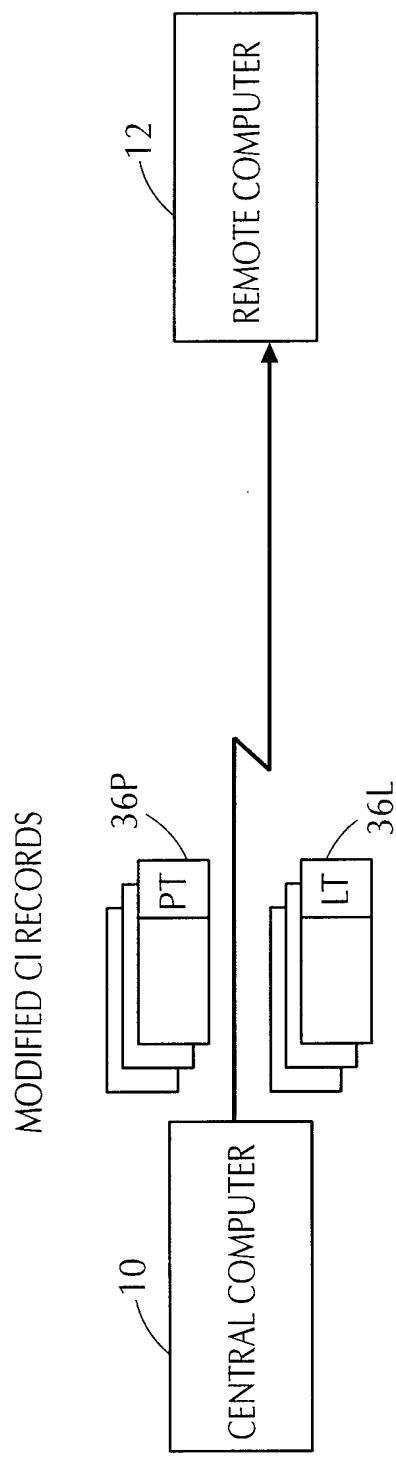
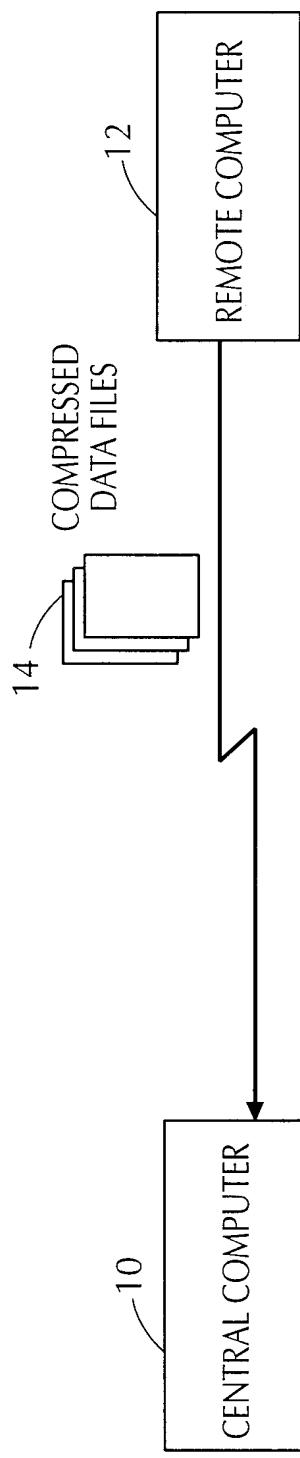
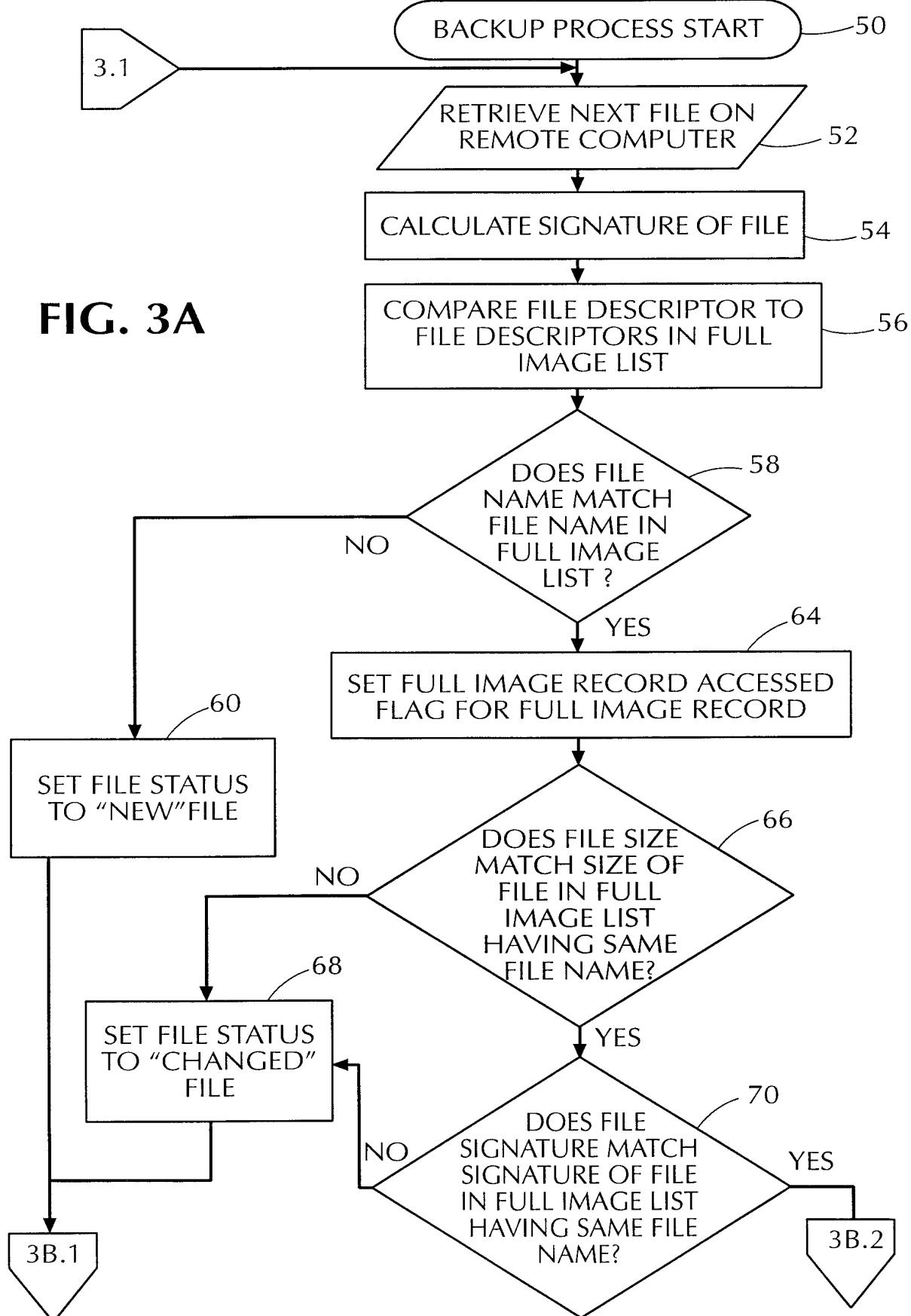


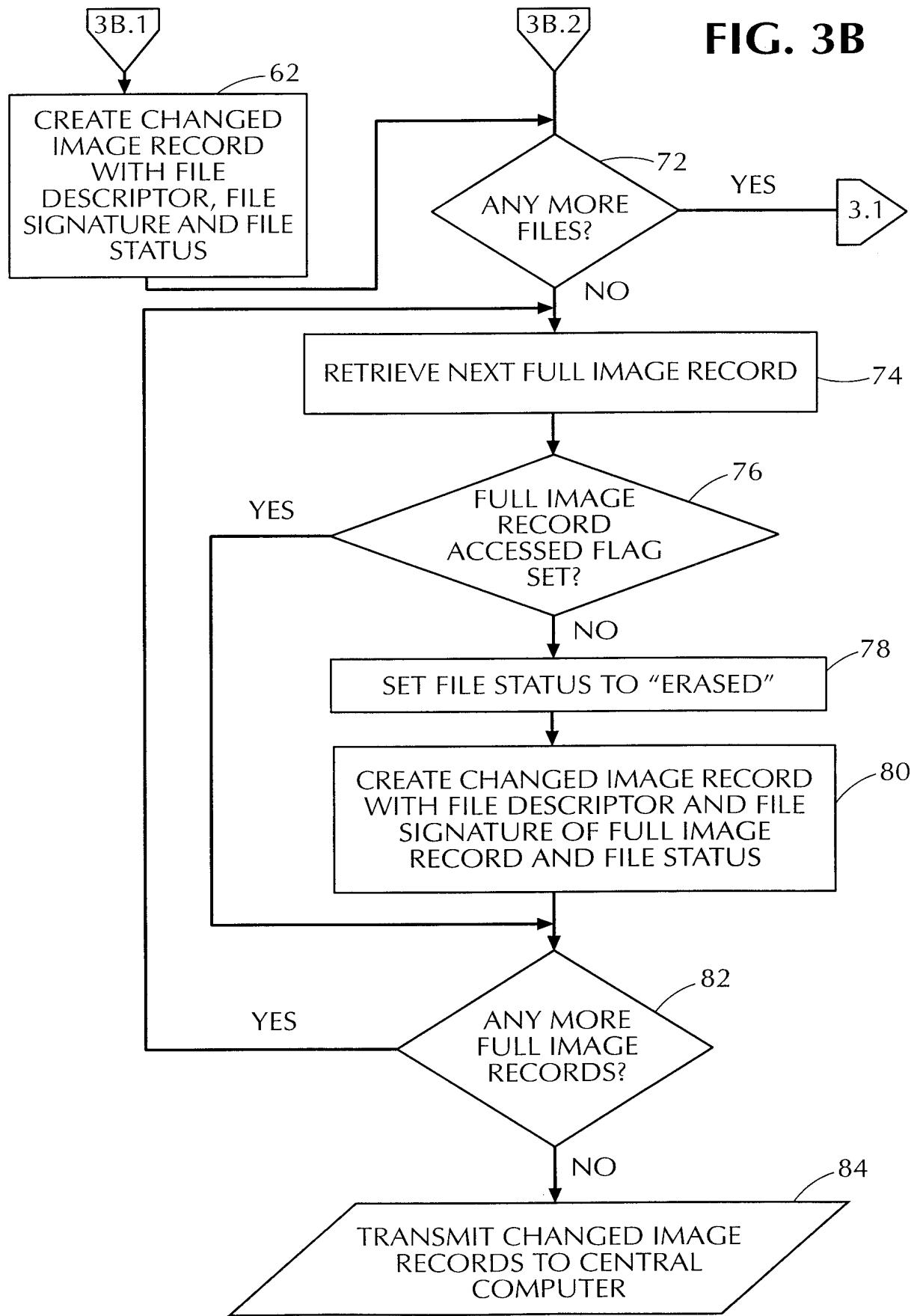
FIG.

**FIG. 2A****FIG. 2B****FIG. 2C**

3 / 20

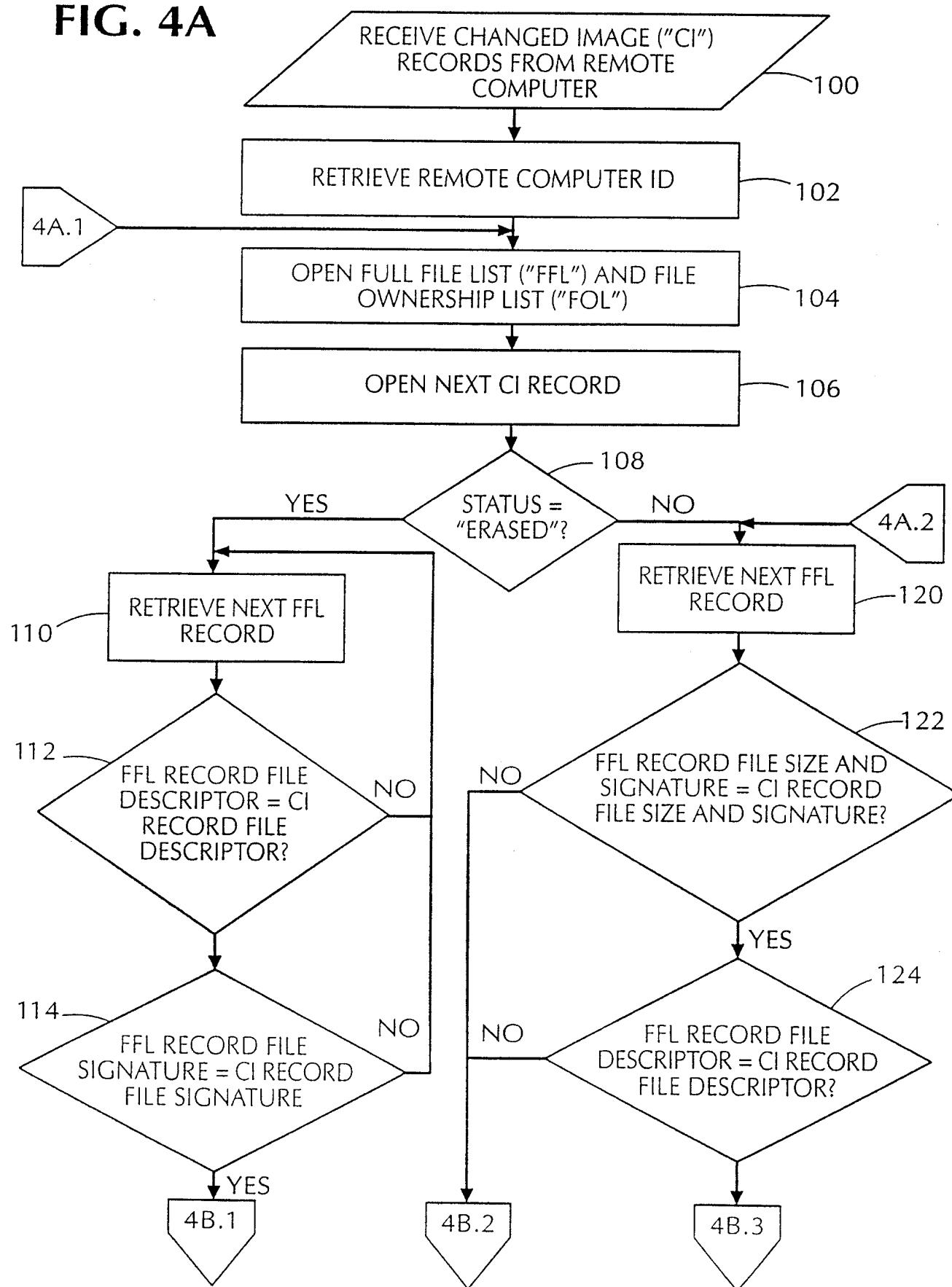


4 / 20



5 / 20

FIG. 4A



6 / 20

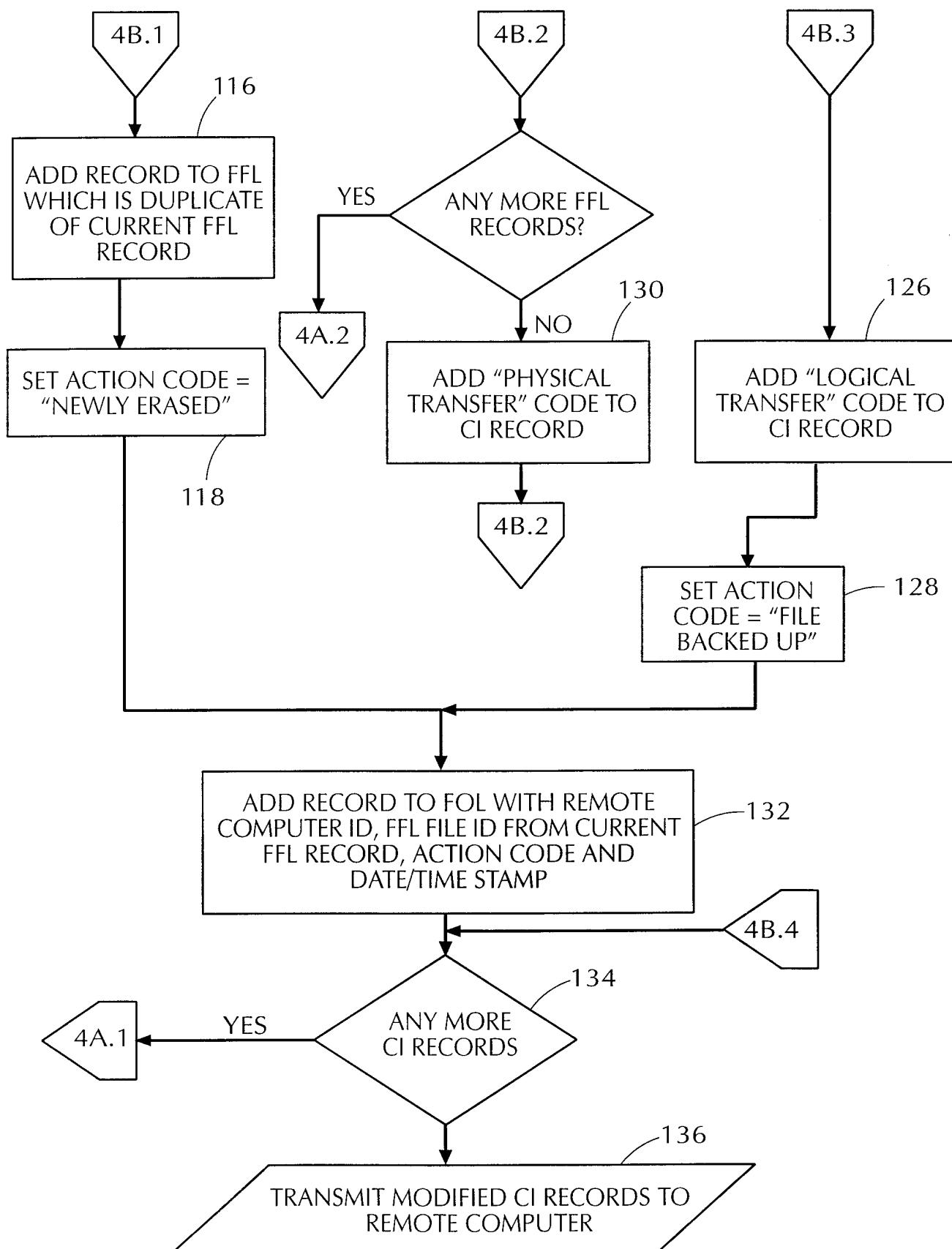


FIG. 4B
SUBSTITUTE SHEET (RULE 26)

7 / 20

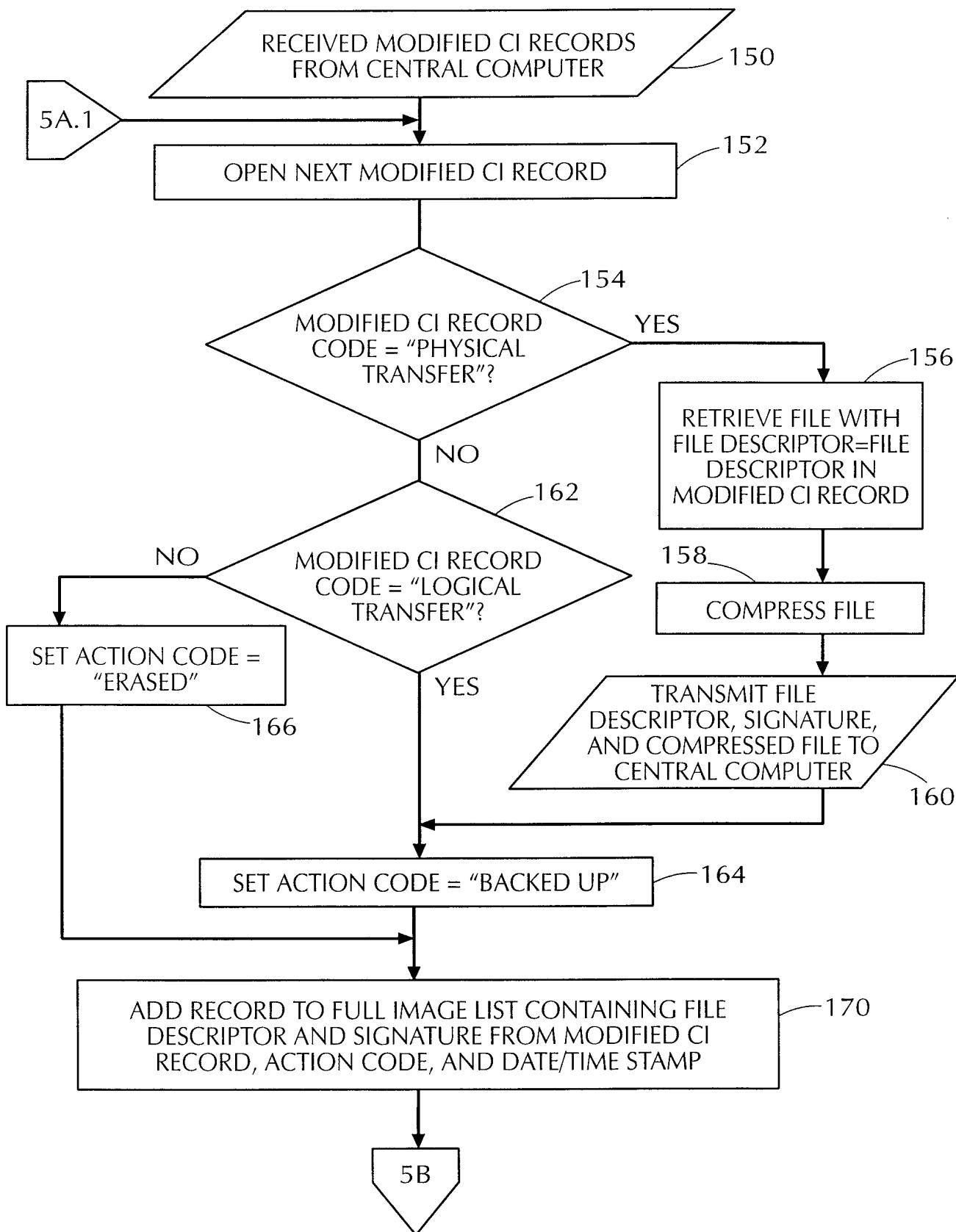
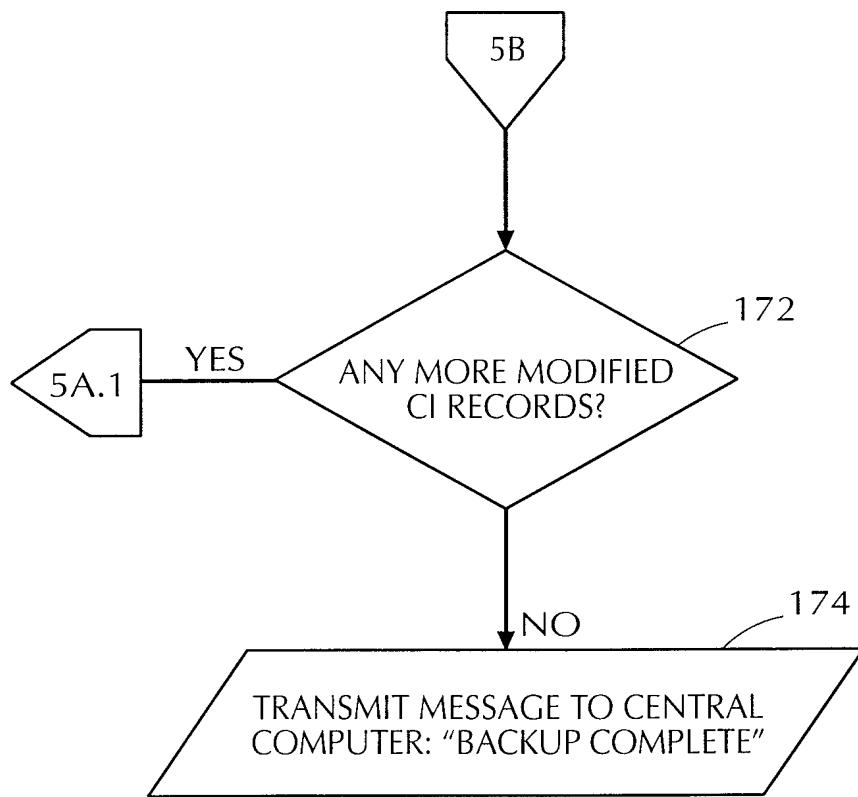
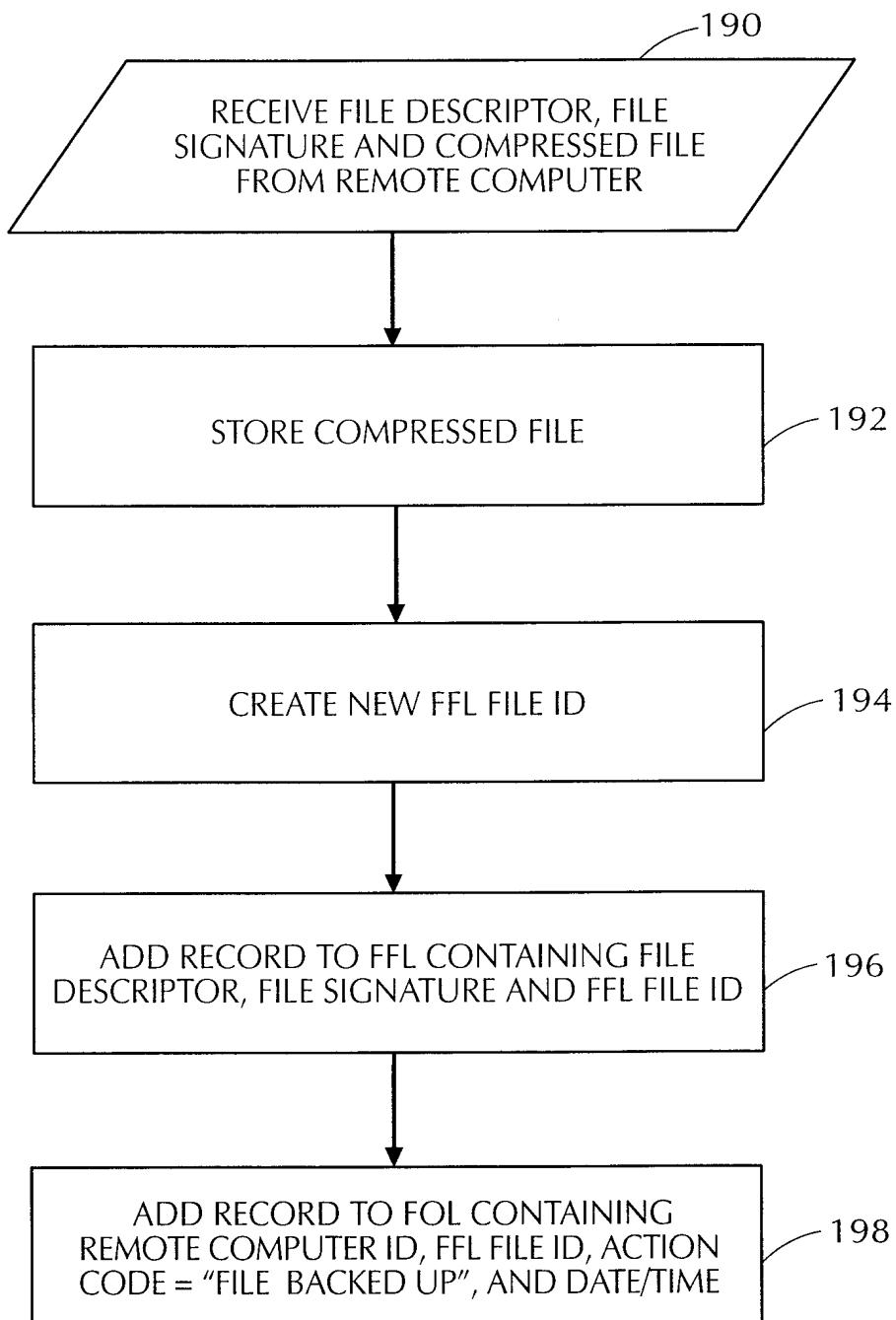


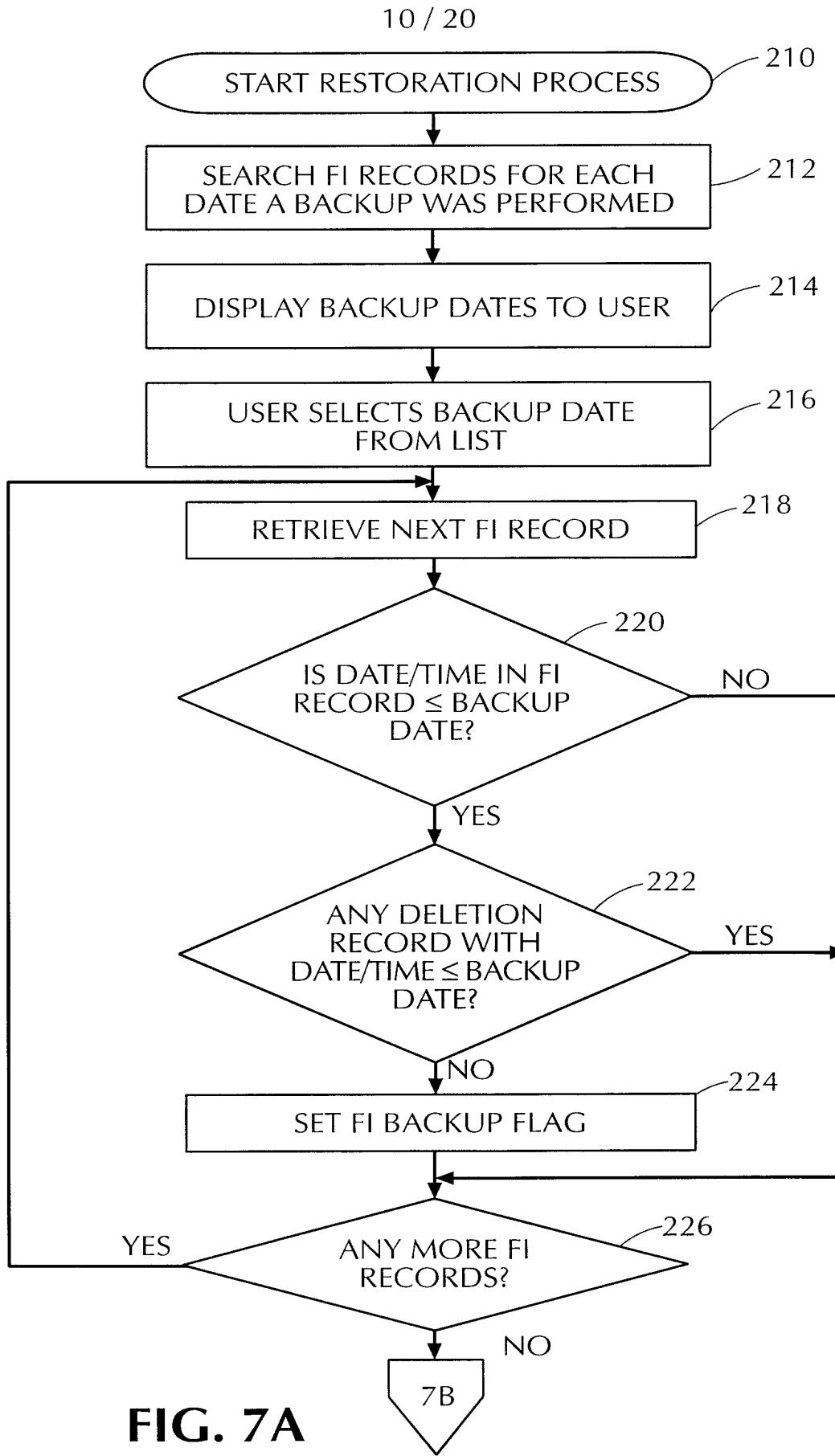
FIG. 5A
SUBSTITUTE SHEET (RULE 26)

8 / 20

**FIG. 5B**

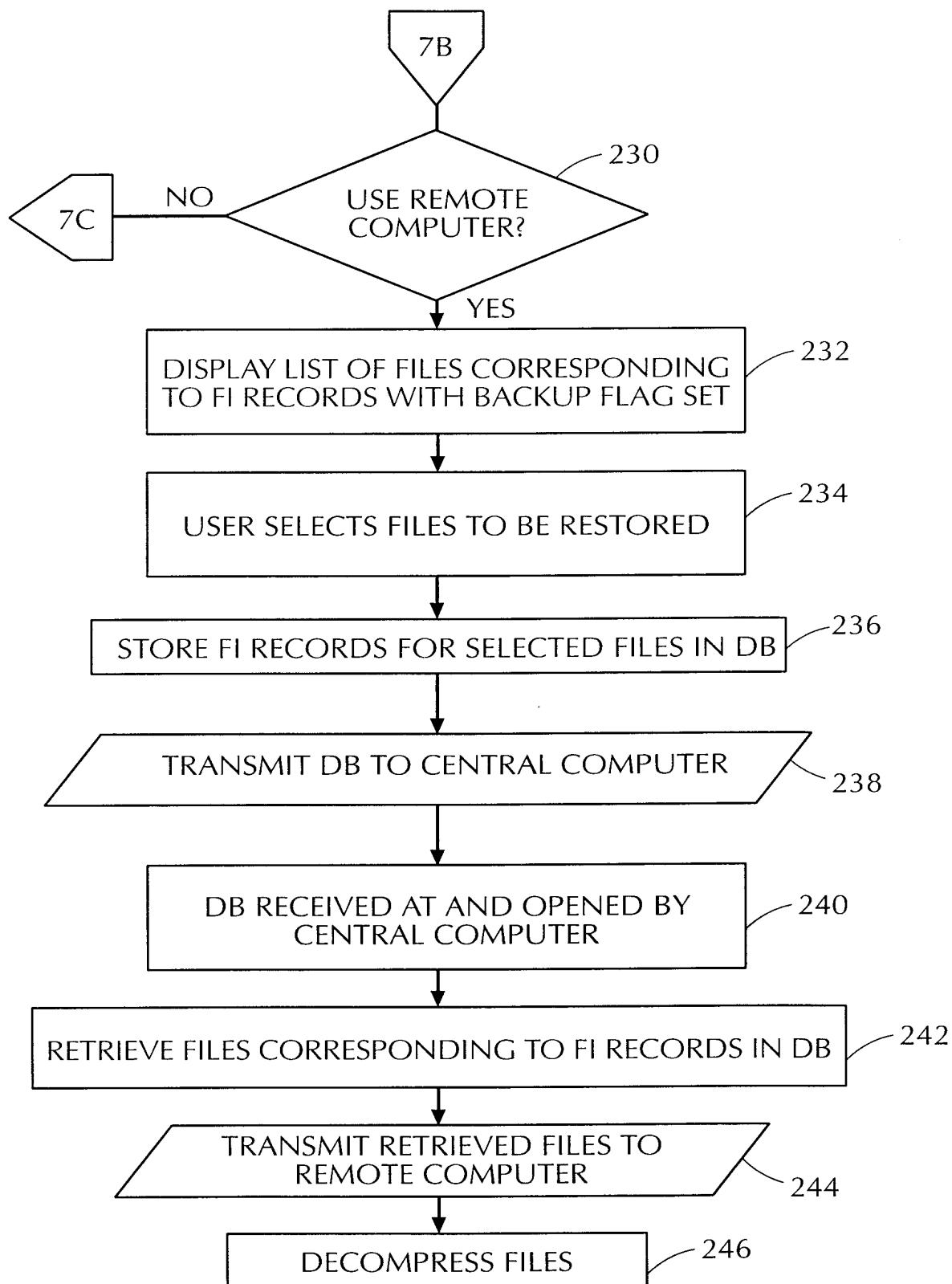
9 / 20

**FIG. 6**

**FIG. 7A**

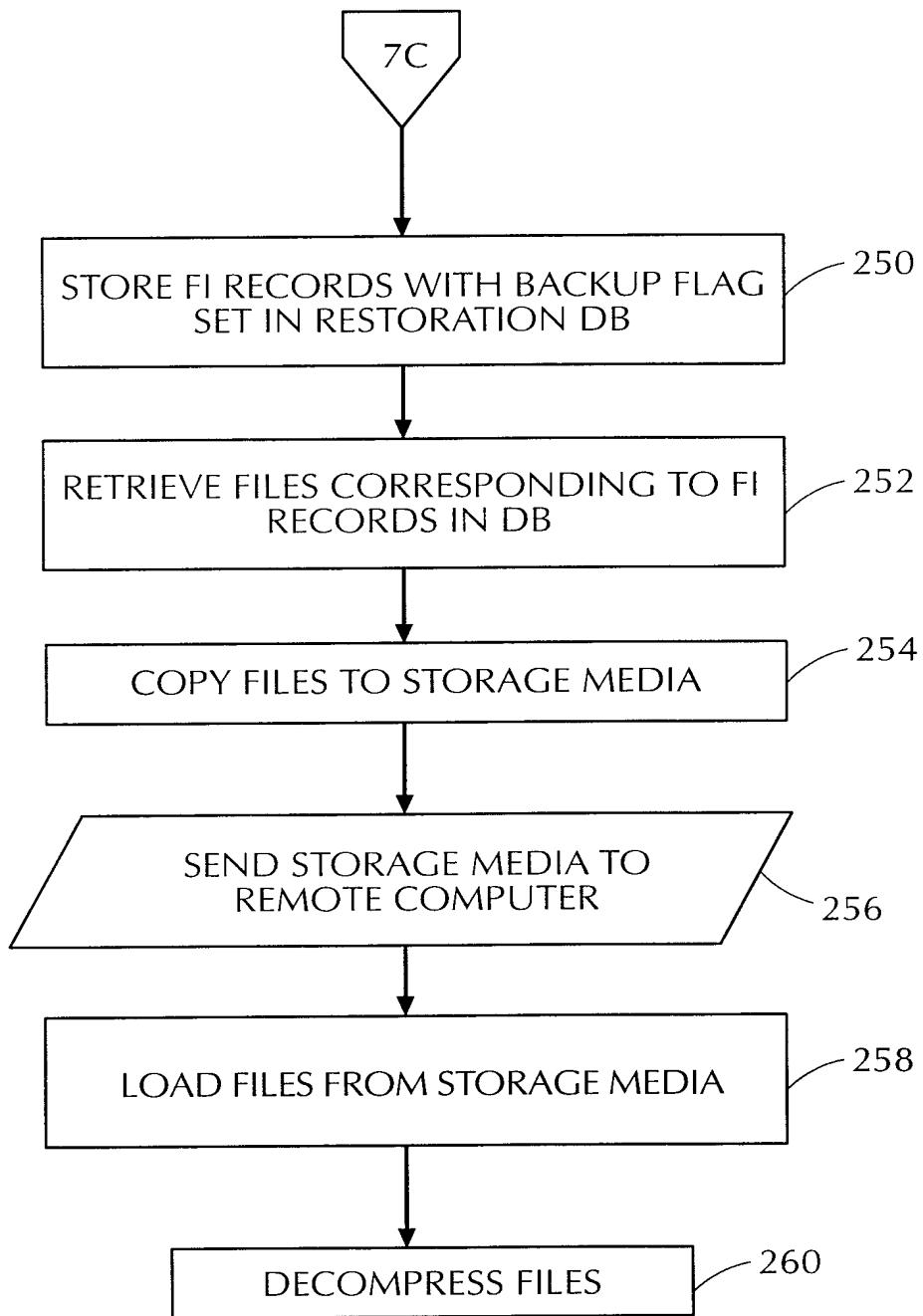
SUBSTITUTE SHEET (RULE 26)

11 / 20

**FIG. 7B**

SUBSTITUTE SHEET (RULE 26)

12 / 20

**FIG. 7C**

SUBSTITUTE SHEET (RULE 26)

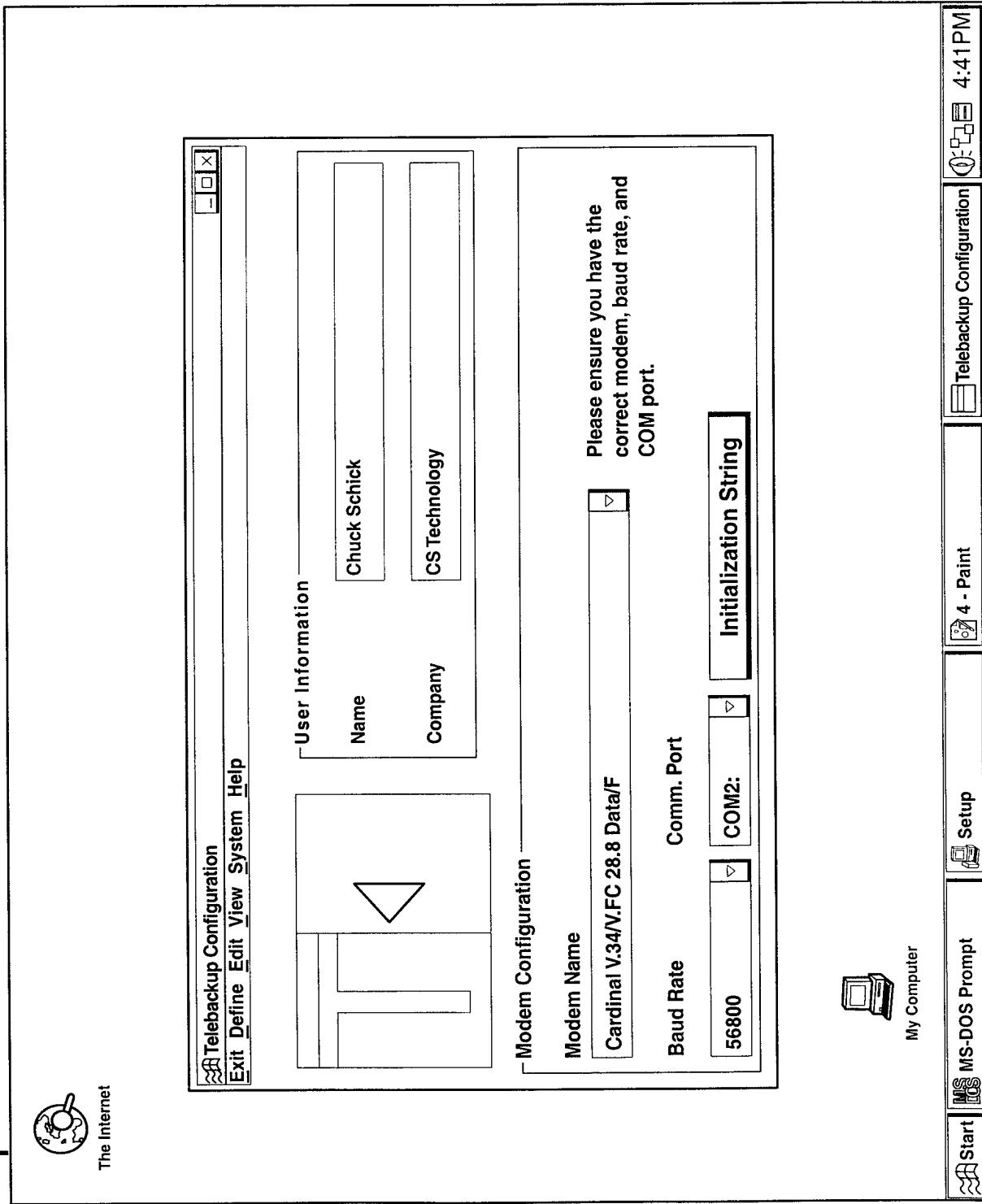
13 / 20

FIG. 8

Step. One



The Internet



[return to previous page](#)

FIG. 9

Step.Two

The Internet

 MS-DOS Prompt

 Setup

 Billing

 4 - Paint

 4:41PM

 My Computer

 Start

Billing

Help

Name: Chuck Schick

Company: CS Technology

Address1: Suite 400

Address2: 26 Fairway Road

Prov/State: R.I.

City: Providence

Phone: 555-5555

Fax: 555-5555

Post/Zip: 90219

Credit Card Information

Name As It Appears on Cards: Charles M. Schick

Number: 0000 0000 0000 0000

Expiry Date: MM YY
12 99

Type: M/C

Purchase Orders

PO#:

Payment Options

Credit Card

Purchase Order

Send Bill

Billing Schedule

Schedule: Monthly
Quarterly
Semi Annual
Annual

Substitute Sheet (Rule 26)

(return to previous page)

FIG. 10

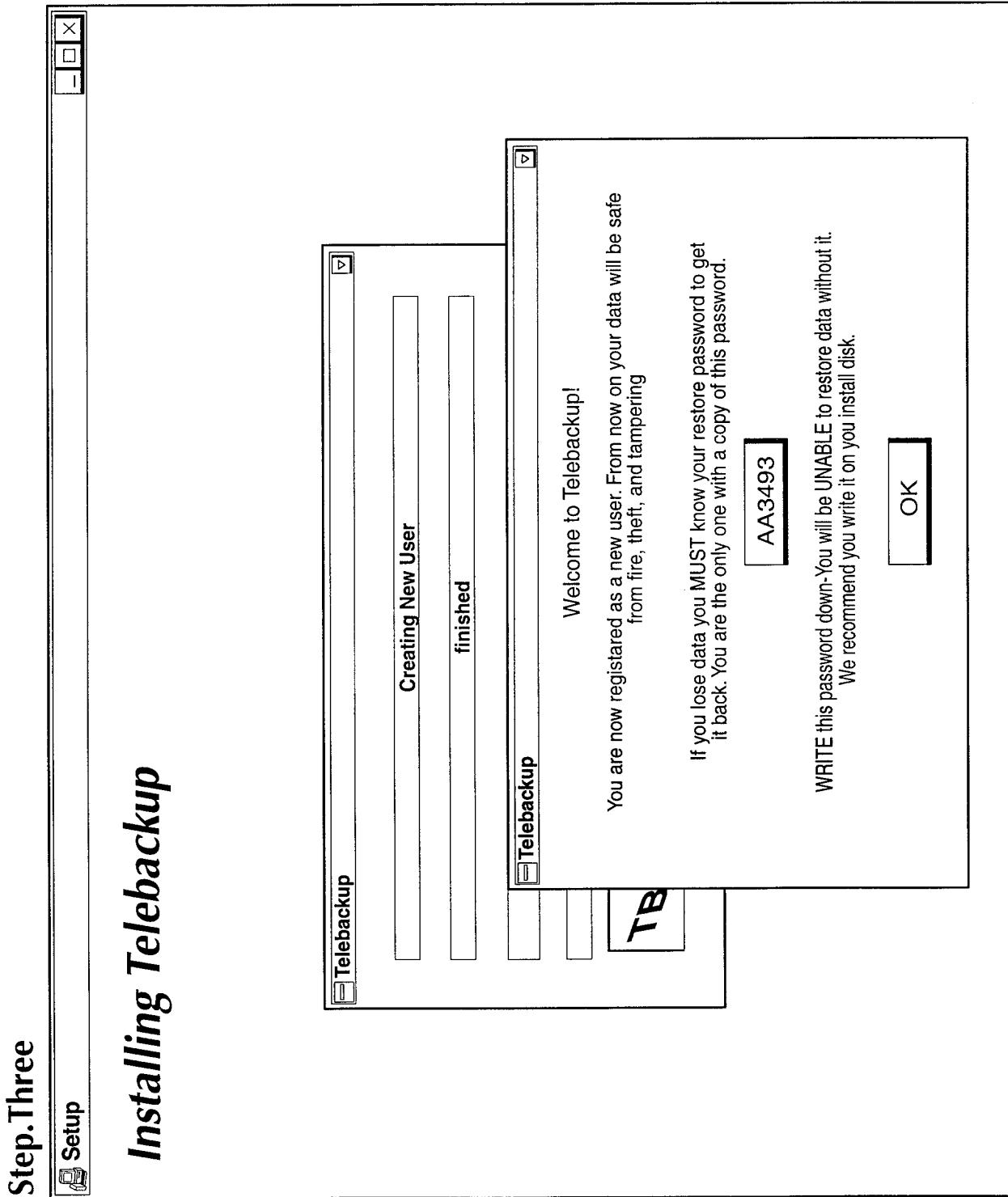
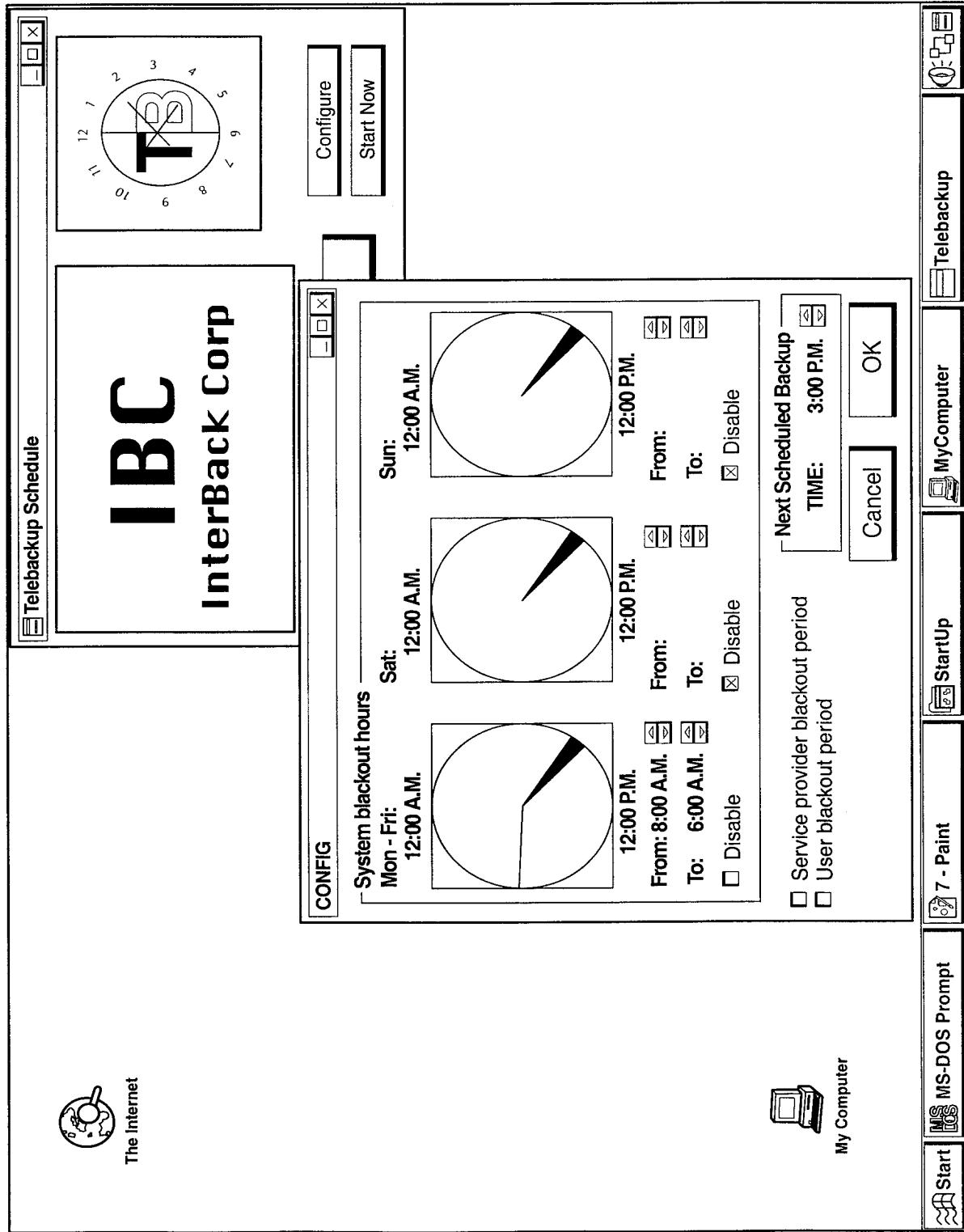


FIG. 11

Step.Four



Step.Five

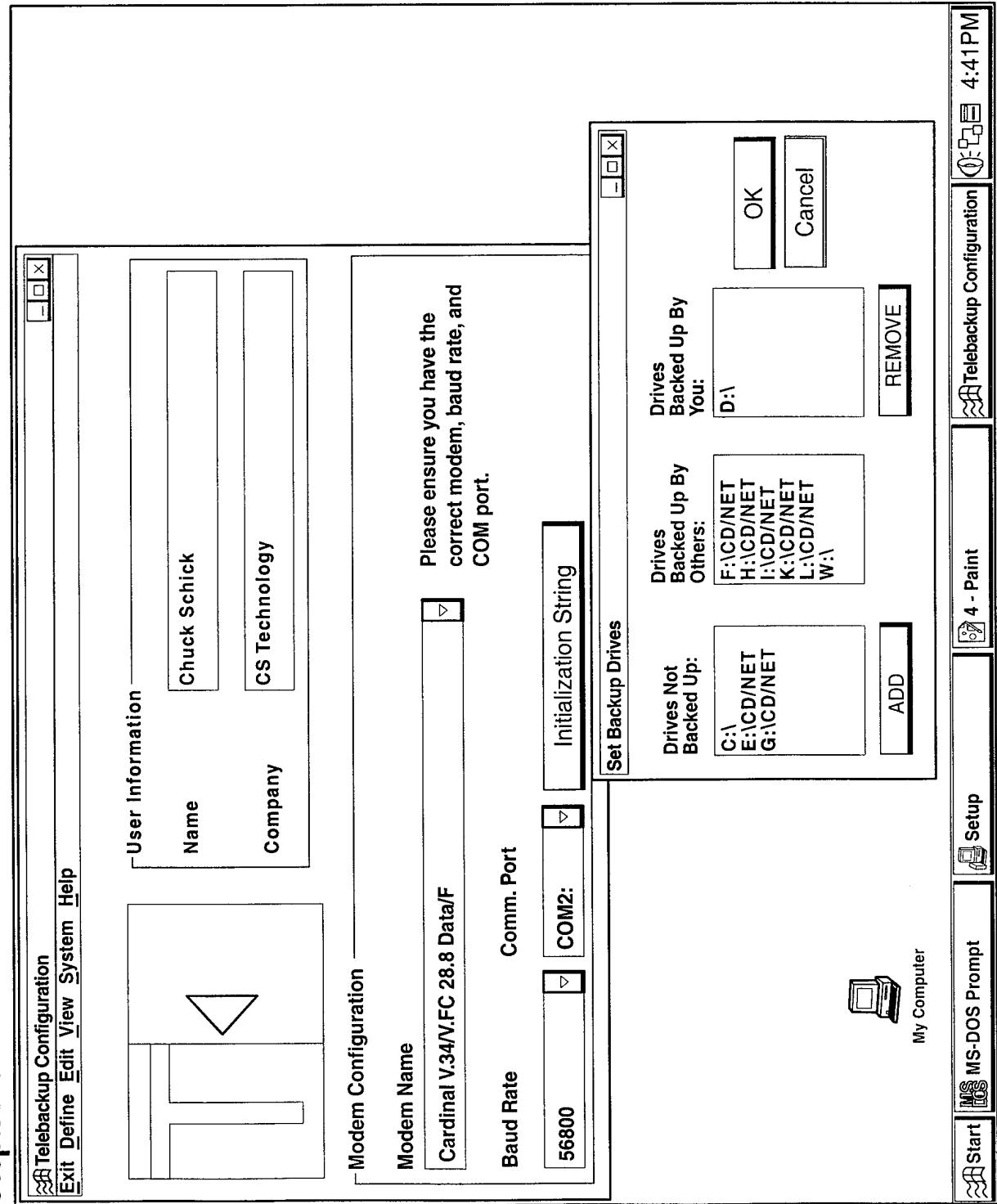


FIG. 12

[\(return to previous page\)](#)

FIG. 13

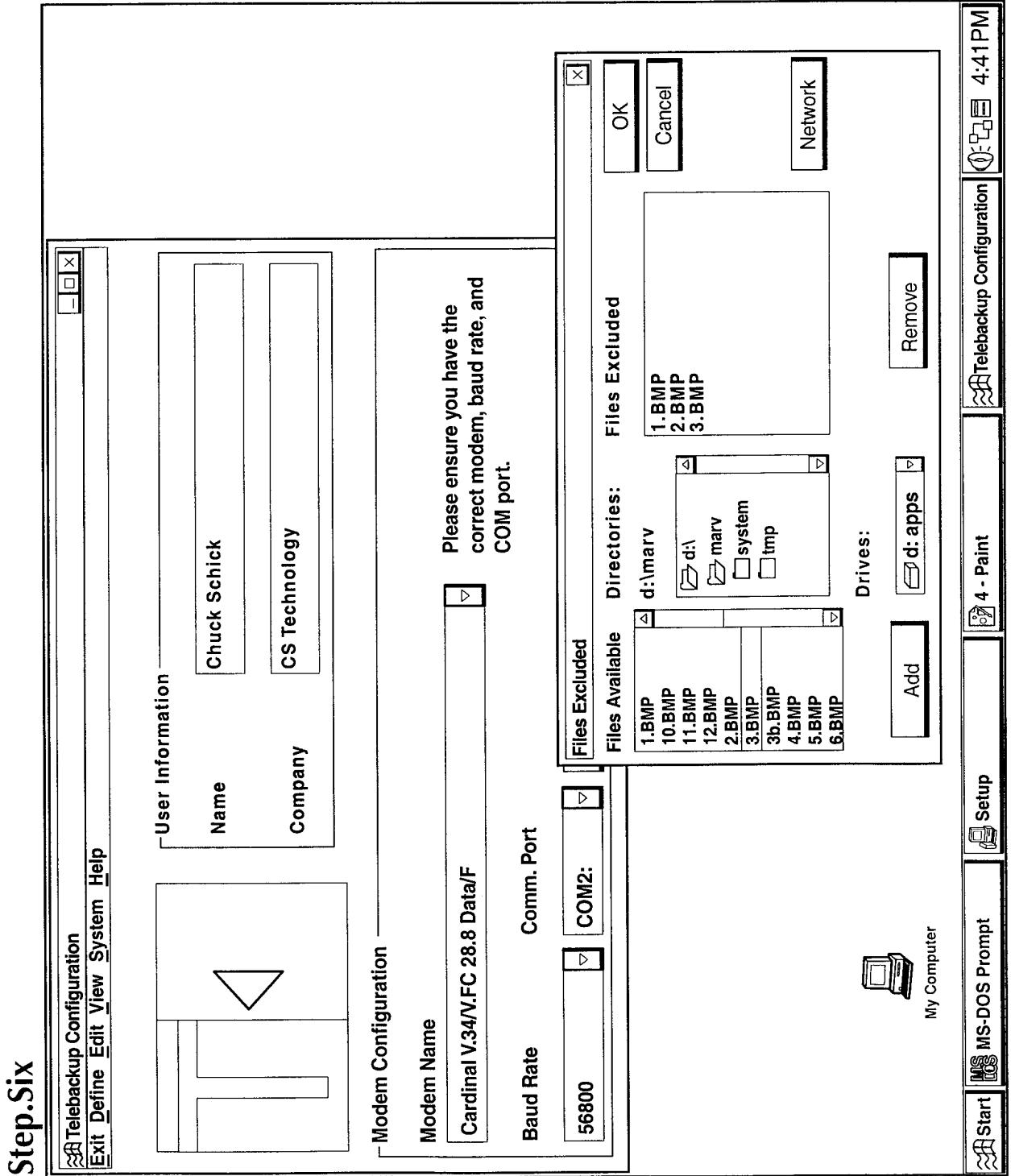
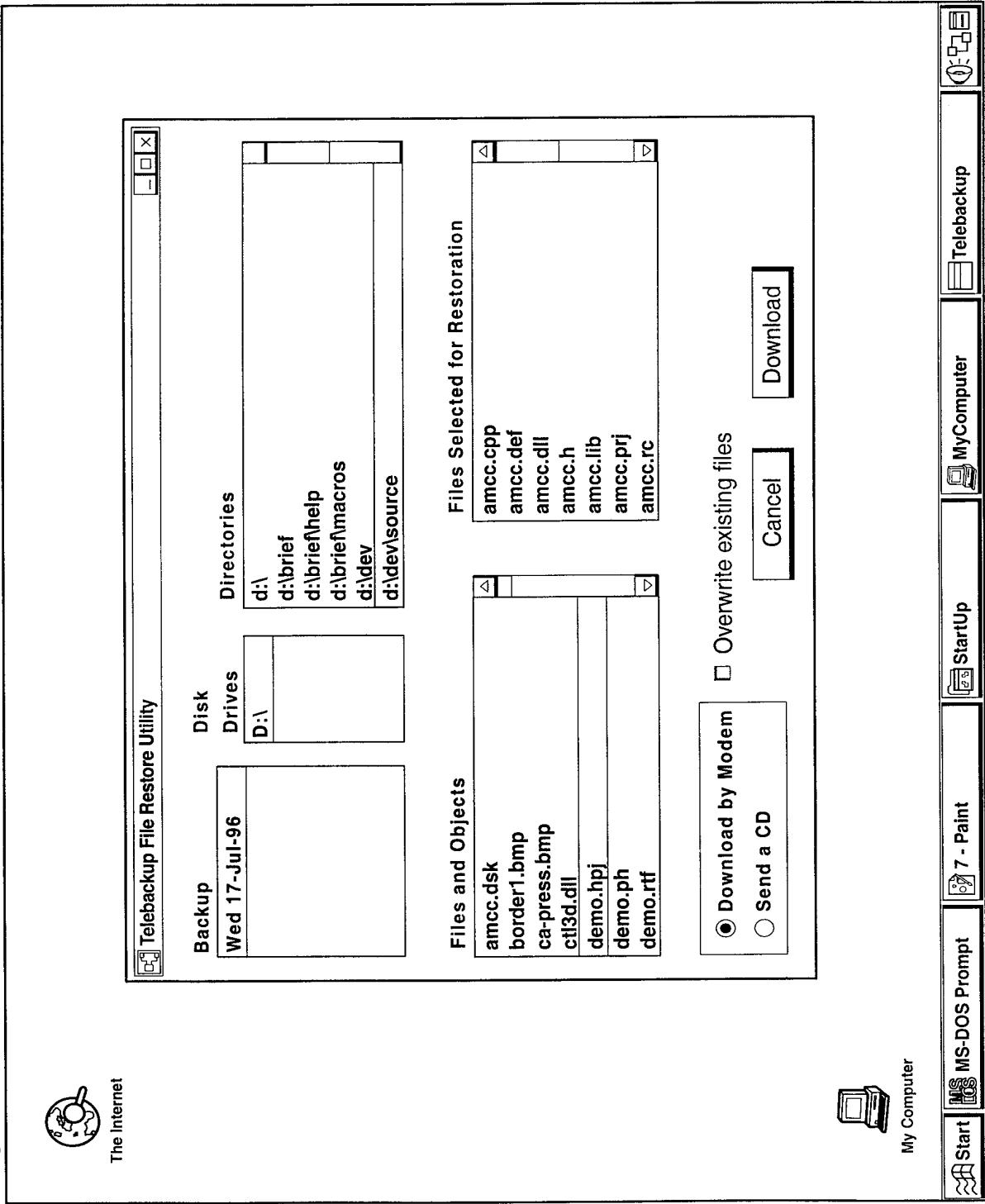


FIG. 14

Step.Seven



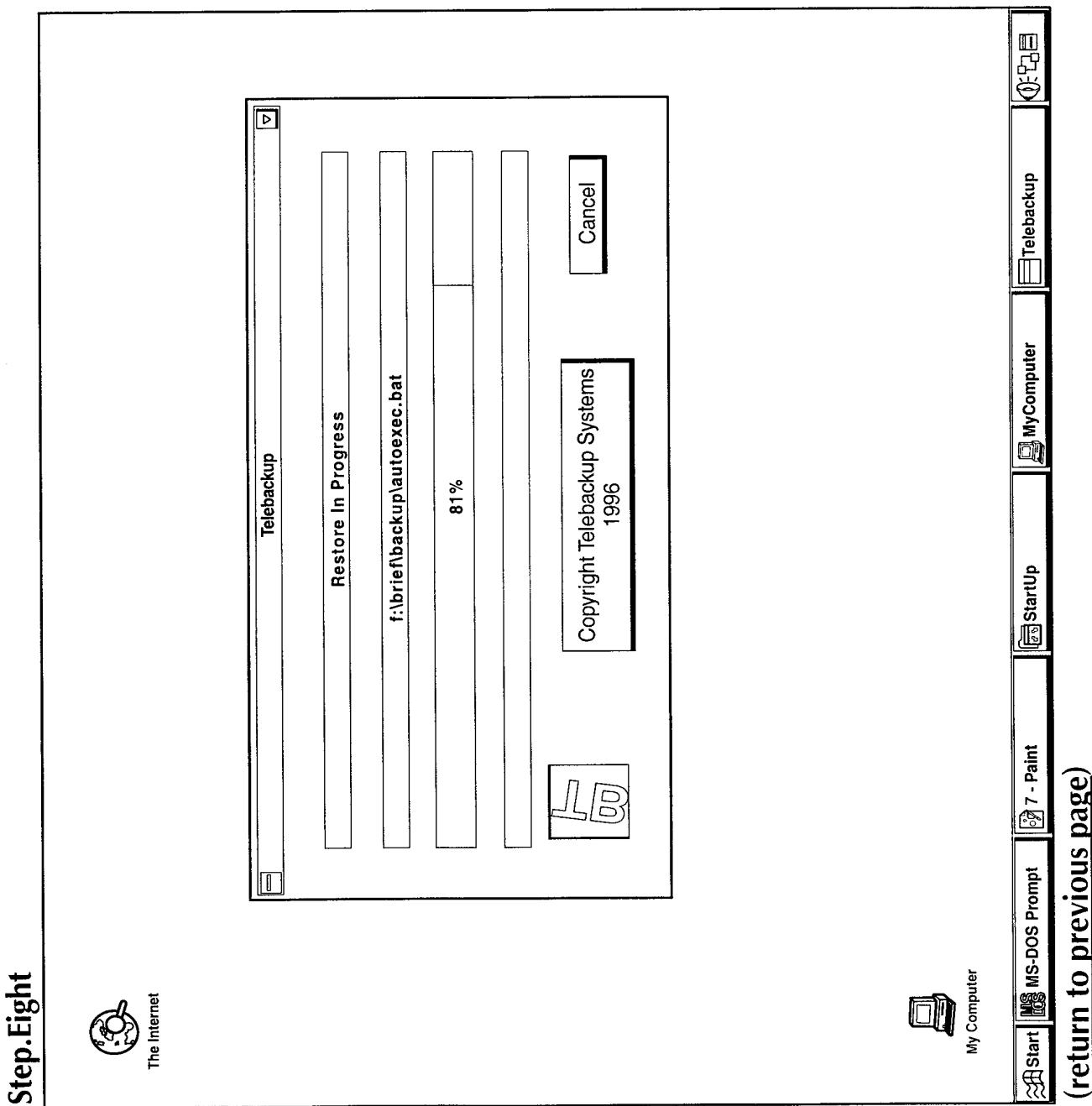
The Internet



[return to previous page](#)

20 / 20

FIG. 15



INTERNATIONAL SEARCH REPORT

Int. Appl. No

PCT/IB 98/01203

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G06F11/14

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	EP 0 774 715 A (STAC ELECTRONICS) 21 May 1997 see column 5, line 56 – column 6, line 38 ---	1-18
Y	EP 0 541 281 A (AMERICAN TELEPHONE AND TELEGRAPH COMPANY) 12 May 1993 see abstract -----	1-18

Further documents are listed in the continuation of box C.

Patent family members are listed in annex.

° Special categories of cited documents :

- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier document but published on or after the international filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search

Date of mailing of the international search report

15 December 1998

23/12/1998

Name and mailing address of the ISA
European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Corremans, G

INTERNATIONAL SEARCH REPORT**Information on patent family members**

International Application No

PCT/IB 98/01203

Patent document cited in search report	Publication date	Patent family member(s)		Publication date
EP 774715	A 21-05-1997	US 5778395 A	JP 10049416 A	07-07-1998 20-02-1998
EP 541281	A 12-05-1993	JP 6290092 A	US 5559991 A	18-10-1994 24-09-1996